12-2015

Exploiting Cross Domain Relationships for Target Recognition

Wei Wang
University of Tennessee - Knoxville, wwang34@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Wei Wang entitled "Exploiting Cross Domain Relationships for Target Recognition." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Computer Engineering.

Hairong Qi, Major Professor

We have read this dissertation and recommend its acceptance:

Jens Gregor, Mark Dean, Russell Zaretzki

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Exploiting Cross Domain Relationships for Target Recognition

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Wei Wang

December 2015

To my parents and my wife.

Acknowledgements

I would like to express my thanks to all the individuals who have inspired, encouraged, and advised me in the past 5 years.

First, I would like to thank my advisor, Dr. Hairong Qi. In research, her courage and insight on choosing research topic, diligence and persistence on doing research, broad knowledge and open minds on solving problems, always guide me to do better work and to be a better researcher. In addition, her cheerful, optimistic and tolerant personality, also teach and infect me to behave better in face of challenges and difficulties in my life. I appreciate all these valuable experiences I learned from her. Meanwhile, I would like to thank my committee professors, Dr. Jens Gregor, Dr. Mark Dean, Dr. Russell Zaretzki, for their important advices and suggestions in my research. I greatly appreciate their time and input to this dissertation.

Second, I also want to thank all my lab-mates for their great help and support in my study and life, including Yang Bai, Sangwoo Moon, Mahmut Karakaya, Li He, Zhibo Wang, Jiajia Luo, Shuangjiang Li, Rui Guo, Liu Liu, Yang Song, Zhifei Zhang, Alireza Rahimpour, Ali Taalimi, Austin Albright, Daniel Capilla, Bryan Bodkin. I really value the friendship we built in AICIP years.

Last but not least, I would like to express my deepest appreciation to my parents and my wife, for their unconditional support and encouragement. Their dedication and love are always the biggest motivation for any of my achievement!

Abstract

Cross domain recognition extracts knowledge from one domain to recognize samples from another domain of interest. The key to solving problems under this umbrella is to find out the latent connections between different domains. In this dissertation, three different cross domain recognition problems are studied by exploiting the relationships between different domains explicitly according to the specific real problems.

First, the problem of cross view action recognition is studied. The same action might seem quite different when observed from different viewpoints. Thus, how to use the training samples from a given camera view and perform recognition in another new view is the key point. In this work, reconstructable paths between different views are built to mirror labeled actions from one source view into one another target view for learning an adaptable classifier. The path learning takes advantage of the joint dictionary learning techniques with exploiting hidden information in the seemingly useless samples, making the recognition performance robust and effective.

Second, the problem of person re-identification is studied, which tries to match pedestrian images in non-overlapping camera views based on appearance features. In this work, we propose to learn a random kernel forest to discriminatively assign a specific distance metric to each pair of local patches from the two images in matching. The forest is composed by multiple decision trees, which are designed to partition the overall space of local patch-pairs into substantial subspaces, where a simple but effective local metric kernel can be defined to minimize the distance of true matches.

Third, the problem of multi-event detection and recognition in smart grid is studied. The signal of multi-event might not be a straightforward combination of some single-event signals because of the correlation among devices. In this work, a concept of "root-pattern" is proposed that can be extracted from a collection of single-event signals, but also transferable to analyse the constituent components of multi-cascading-event signals based on an over-complete dictionary, which is designed according to the "root-patterns" with temporal information subtly embedded.

The correctness and effectiveness of the proposed approaches have been evaluated by extensive experiments.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Cross Domain Recognition

Massive data is generated from various areas in this data-centric era. For example, the users of Facebook, Twitter, Youtube, etc, contribute incredible amount of data every data; the cameras for city security surveillance also generate a huge volume of monitoring video sequences every minute; research in both science and engineering also collect a large amount of observational or synthetic data. Therefore, it is essential to analyse data from multiple sources collaboratively to extract information and make new discovery. Meanwhile, the analysis of the multi-source data also poses a great challenge as the data maybe generated with non-identical attributes or distribution.

Machine learning has been thoroughly investigated for decades and widely applied to many areas, such as data mining (Witten et al., 2011), computer vision (Bishop, 2007). Traditional machine learning usually assumes the data property is relatively stable across the learning and recognition phases, where the data property refers to the samples feature vector or the data distribution, such that the classification model learned from training data can be used for recognition of the testing data. However, this assumption is usually violated in many real world problems where either the

attributes of the data instances or the structure of data distribution varies across the training and testing data.

Formally, suppose $(x, y)_{train}$ and $(x, y)_{test}$ are samples drawn from certain training distribution D_{train} and testing distribution D_{test}, respectively, where x denotes the instance feature vector and y denotes the class label. Traditional machine learning usually assumes the feature representations x share a common feature space, i.e., the same attributes of the features, meanwhile, $P_{train}(y|x) = P_{test}(y|x)$, meaning the training data and testing data are under the same distribution, i.e., $D_{train} = D_{test}$. However, if we permit either assumption being relaxed, i.e., allow the attributes of instance features or the data distribution vary across the training and testing data, the performance of traditional machine learning and recognition approaches will degrade a lot. In other words, these approaches are not able to adjust themselves in recognition of the testing samples with inconsistent data property, since the prediction from the classification models based on these approaches become uncertain.

Traditional machine learning usually requires access of sufficient labeled training data to learn robust models or classifiers for the purpose of better prediction on the unseen testing data (Vapnik, 1998; Hastie et al., 2009). However, if the testing data is not from an identical or similar domain of the training data D_{train}, it usually requires large human effort to obtain labeled training data under the same data distribution D_{test}. From the perspective of human psychology, the model learned in D_{train} is still useful if we can find certain latent connection between D_{train} and D_{test}, and then the models can be modified to effectively recognize testing data. Therefore, how to make use of the knowledge extracted from training data via the latent connection between different domains presents a challenging problem.

In recent decades, cross domain recognition becomes a growing research area with a wide range of applications. For example, the video concept detection (Duan et al., 2011). It extracts semantic concepts, such as "person", "animal", "building" and so on, as classification models from video data in some source domain, then uses these models to detect the concepts in other domains. Here, a domain is a TV channel, such

as CCTV, CBS, CNN, and NBC. For instance, the TRECVID dataset (Smeaton and Over, 2003) is a typical multi-domain video collection which has news videos from different TV channels. As a result, even for the same concept, the data distribution of CCTV is usually different from that of CNN. Another example is for web-page categorization (Zhuang et al., 2010). One typical application is using classification model to retrieve course main pages from all the web-pages in an university website. We may create training data by manually labeling a collection of main course pages from the university website with a lot of human efforts, but an alternative way is to use some already labeled main course pages from other universities as the training data. However, different universities have different templates for course pages, where the terms used also may be different. As part of human nature, we prefer to investigate information gained from previous efforts, instead of restart a new endeaver, to solve the similar tasks. Therefore, we desire to transfer the knowledge from one domain into another one via certain guidelines, e.g., the semantic concept from CCTV to CNN, or the course page style from one university to another, such that we do not need to learn models for each individual data domain while targeting one common recognition purpose. Except for the aforementioned real-world applications, there are also many other applications involving the same cross domain recognition issue, such as image classification (Liu et al., 2011b; Shekhar et al., 2013; Fernando et al., 2013), image clustering (Yang et al., 2009; Gopalan, 2013), natural language processing (Blitzer et al., 2006; Arnold et al., 2007; Wu et al., 2009), wireless sensor networks (Pan et al., 2008, 2011; Yin et al., 2008), sentiment classification (Li et al., 2009a; Remus, 2012), and so on.

1.2 Motivations

In order to remove the constraint of same-data-distribution in traditional machine learning approaches, *knowledge transfer* has been proposed by allowing the training and testing data to be from different yet implicitly correlated domains. The formal

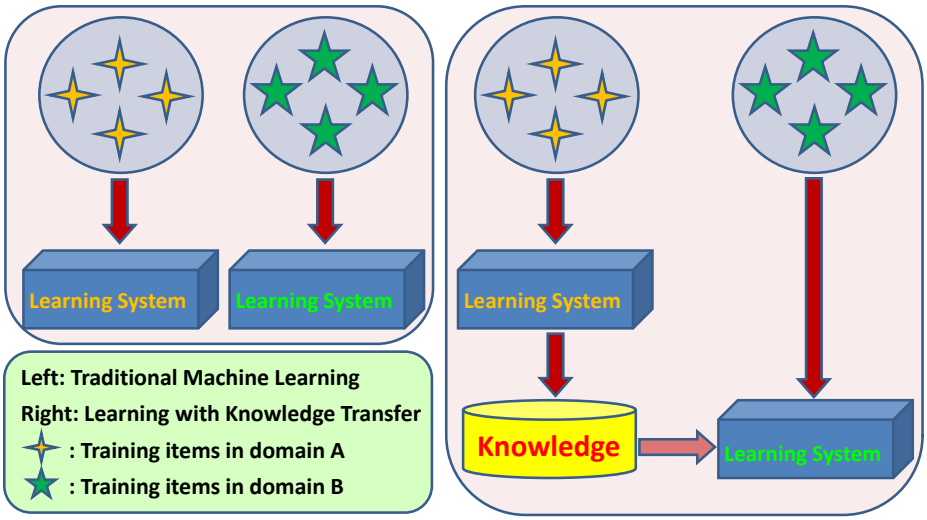

Figure 1.1: A comparison between traditional machine learning and learning with knowledge transfer.

definition of knowledge transfer is to apply previous knowledge extracted from one or multiple existing domains/tasks to improve the learning in the new domains/tasks of interest. Nowadays, knowledge transfer serves as a general term of machine learning that covers a variety of approaches including multi-task learning (Harpale and Yang, 2010), domain adaptation (Jhuo et al., 2012), sample selection bias (Pan and Yang, 2010), covariate shift (Bruzzone and Marconcini, 2013), etc. Figure 1.1 provides an intuitive illustration describing the difference between traditional machine learning and cross domain knowledge transfer. From the illustration, we can observe that the key to address real-world cross domain recognition problems that previously hard to solve is how to exploit the latent relationships between the different data domains.

Therefore, the main purpose of this dissertation is to tackle the problems under the umbrella that the training data and testing data are in different domains. Specifically, we exploit different latent relationships across data domains to solve several real-world cross domain target recognition problems:

First, **action recognition across camera views:** action recognition is essential to many real world applications, such as visual surveillance, video retrieval, human-computer interaction, etc. The spatio-temporal features are popularly used as action

representation in typical action recognition settings. However, these features are view-dependent. Although they are powerful in discriminating actions observed from similar viewpoints, the same action may look quite different if viewed from different camera viewpoints. Due to the poor generalization of these view-dependent features across different camera views, the performance of these features degrades significantly when the observing viewpoint changes. This also can be verified by the fact that the magnitude of inter-class variation of action characteristics, which distinguishes one action from the others in the same view, may be even smaller than the intra-class variation caused by the change of viewpoints. In this problem, we assume to be given pairwise learning samples from two cameras that recorded a group of actions performed by different subjects, while the labeled training data for classification model only exist in one source data domain. The objective is to enable the actions observed in a target view to be recognizable by a classifier trained by labeled samples observed in a source view. Thus, how to exploit the latent relationship for actions in different views is the key for solving this problem.

Second, **person re-identification across non-overlapping cameras:** person re-identification is to discover correct matches of pedestrian images observed in non-overlapping camera views by visual features. It is able to save human effort by avoid exhausting search of an interested person from large amount of video sequences, and has attracted considerable attentions, particularly in the surveillance community for its importance in pedestrian retrieval, event detection, and multi-camera tracking. After years of research, this problem is still extremely challenging, and its difficulty mainly attributes to the significant disjoint of the non-overlapping cameras. A person observed in different camera views often suffers from changes in viewpoints, poses, illuminations, complex backgrounds and occlusions. Meanwhile, different people also might share similar appearance. All the factors make two images of the same person look different while images of different people look similar. Therefore, direct matching of the visual features of person images from different cameras is not reliable for the challenges. In this problem, we also assume to be given training samples in pair from

two cameras that recorded a group of individual persons, while the objective is to identify for a query image recorded in one camera from a large number of candidates in the gallery of another camera. Thus, how to exploit the latent relationship between the two cameras is the key for solving this problem.

Third, **multi-event detection and recognition in smart grid:** Event analysis has been an important component in any situational awareness systems, i.e., smart power grid system. When an event occurs in a smart grid, the imbalance between generation and load consumption causes sudden frequency changes within the system that can also be used as an indicator for event disturbance. Although successful, the state-of-the-art techniques can only handle disturbances caused by a single event. If multiple cascading events are involved, existing techniques can only detect frequency disturbances caused by the initial one, and the frequency disturbances from successive events might be overshadowed by the continued frequency fluctuation from the initial event. Thus, how to determine the number of events that occurred and identify the types of events that involved with precise estimation of occurring time using simply the observed 1-D signal is a very challenging problem. In this problem, we are given enough instances of signal observed in single event disturbances, while the objective is to analyse the signal observed in event of multiple cascading disturbances. Thus, how to exploit the physical latent relationship between the given single event signals and the multi-event signals is the key for solving this problem.

1.3 Contributions

In this dissertation, approaches related to how to learn the implicit latent relationships between different data domains are proposed for the aforementioned challenges. In summary, our contributions include:

- For the cross view action recognition problem, a reconstructable path learned from view dependent action representations (RP-VDR) from both cameras is

proposed. The RP-VDR allows labelled training samples from a source domain to be reconstructed into the target domain, such that we are able to train a classifier in the target domain with these mirrored samples from source domain. If multiple source views are available, the mirrored samples from different source domains can also be used together for learning a stronger classifier.

- To exploit the hidden information existing in some of the seemingly useless action samples in each camera view, we also proposed RP-VDRh to facilitate the learning of the reconstructable path, such that the path can achieve better prediction with much less restricted learning samples. Alternate dictionary learning technique is used to realize the path learning, such that the structure information of each view domain can be fully exploited and the discrimination among action categories can be well preserved after reconstruction.

- For the problem of person re-identification across cameras, unlike other existing works to learn a fixed distance metric for the pedestrian images matching, we proposed to learn a random kernel forest that is able to discriminatively assign the optimal local metric kernel to each local region of the query image, such that the distance between images of the same individual captured in different cameras can be better minimized. The metric kernel forest is optimized based on aligned local image patch-pairs. Its behind gist is to decompose the complex inter-camera transformation into a lot of simple local transforms, thus the space of the local patch-pairs is partitioned into many subspaces based on the criteria of consistency patch-to-patch transform via feature split in a decision tree. Multiple trees compose a forest to prevent over-fitting and generate good prediction.

- For the multi-event analysis problem, physical analysis reveals that when the multi-event occurs in a cascading fashion, the measurement taken at sensors would more than likely be a "mixture" of several constituent component signals. We therefore proposed to learn a group of "root-patterns", which are modeled

as latent common patterns for both the domains of single event and multi-event, to represent these constituent component signals in a dictionary. Consider that each constituent event may occur at different time, temporal stamps are also subtly embedded in each column of the dictionary for temporal awareness. By decompose the signal of multi-event according to the constructed dictionary with sparsity and non-negativity constraints, we are able to detect, recognize each constituent event and also identify their occurring times simultaneously in one step with high accuracy.

1.4 Dissertation Organization

The dissertation is organized as follows: Chapter 2 presents a literature review on our studied problems. Chapter 3 studies the problem of action recognition across camera views, as well as the experimental results. Chapter 4 studies the problem of person re-identification in non-overlapped camera networks, as well as the experimental results. In Chapter 5, we study the problem of multi-event detection and recognition in smart grid system, and evaluate experimental results on both synthetic data and real world cases. Finally, conclusion and future research are discussed in Chapter 6.

Chapter 2

Literature Review

2.1 Background

Training on data from task domain might be the most straightforward solution for cross domain recognition problems, since the training data would be drawn from the identical distribution as the data ultimately test on. However, the drawback with this solution is that it usually requires a large amount of labeled data, which are often not available in the task domain. From the perspective of human psychology, the better solution for cross domain recognition is to investigate how to transfer the knowledge learned from previous domain to the new task domain that shares certain kinds of latent statistical connections (Thorndike and Woodworth, 1901; Elli, 1965).

Generally, the most popular approach for solving this kind of problems is learning with knowledge transfer. Under the umbrella of cross domain recognition, there are several sub-problems. The transfer problem is called as "domain adaptation" (Daume and Marcu, 2006) if only the data being analysed is allowed to vary; while the transfer problem is called as "multi-task learning" (Caruana, 1997) if the task being learned is allowed to vary. Both of them are typical transfer learning problems that require an effective way to manipulate the classifier learned in source domain to adapt to the unique property of task domain. For example, the adaptive SVM (Yang et al., 2007)

uses its incremental learning ability to adjust the classifiers $\mathcal{C}^{old}$ learned from source domain to classify testing samples, thus there is no necessity for re-training the entire model using training data from source domain again.

In previous works, different research has made different assumptions about the relationship across data domains. In supervised setting, the comparison of both the marginal and conditional data distributions among different domains is permitted to look for patterns with strong generalizability across data domains (Romero and McCree, 2014; Li et al., 2014a; Ma et al., 2015), or to examine the common structure of correlated problems (Arnold et al., 2008; Chen et al., 2014; Zhang and Mahoor, 2014). Unsupervised (Ni et al., 2013; Baktashmotlagh et al., 2013) and semi-supervised (Cheng and Pan, 2014; Xiao and Guo, 2015) settings are also investigated to quantify these inter-domain relationships. In this dissertation, the study mainly focus on the supervised settings, and the implicit relationships between different data domains are explicitly exploited and modeled. It is also worth mentioning that some researches also use metric learning to bridge the correlation between different domains by mining the commonly shared information (Kulis et al., 2011; Luo et al., 2014b; Ding et al., 2015). Metric learning is to learn a distance metric from a given set of paired samples of similar/dissimilar that preserves the distance relationship in both of the training and testing data. We also investigated this approach in this dissertation.

2.2 Cross View Action Recognition

The purpose of this part of work is to recognize human actions across changes in the observers viewpoint. Opportunities for the use of action analysis are in areas such as surveillance, video indexing/retrieval, and human-computer interaction, etc. To handle this cross-view recognition problem, there generally have been two categories of research directions. One direction relies on extracting effective view-invariant action features (Junejo et al., 2008; Rao et al., 2002). Another emerging family of approaches is based on *transfer learning* that encourages the actions recorded from different views

to be represented by certain commonly shared representations (Farhadi and Tabrizi, 2008; Liu et al., 2011a) to achieve cross view invariance.

2.2.1 View Invariant Action Features

In the first category of research line, which mainly relies on view invariant features extraction, a lot of approaches were proposed. In (Juneio et al., 2008), actions were represented by view stable temporal self-similarity matrices. Alternatively, view-independent can also be achieved by 3D models. For example, (Weinland et al., 2007) described actions by 3D examplars and performed recognition via matching in projected 2D space. In (Li et al., 2007), 3D shapes and poses were directly estimated from multiple-view inputs for action recognition. Multiple-camera systems were also investigated (Luo et al., 2013a). (Yilmaz and Shah, 2005) proposed to exploit dynamic epipolar geometry by imposing temporal fundamental matrix and (Paramesmaran and Chellappa, 2006) exploited projective invariants of coplanar landmark points for view-invariant feature extraction. Inferable classifiers were also proposed in (Weinland et al., 2010; Wu and Jia, 2012), where the former handled view changes by learning a classifier based on examples taken from various views and the latter proposed to learn a kernelized structural SVM which regards the view label of action as a latent variable and implicitly infer it during both learning and inference. Nevertheless, the biggest limitation for these techniques is the relatively complicated reasoning of view alignment.

2.2.2 Cross View Knowledge Transfer

Transfer learning applies knowledge learned in one task to novel tasks or new domains sharing some commonality. It has been explored in many applications, such as image classification and super-resolution, visual domain adaptation (Quattoni et al., 2008; Wang et al., 2012; Jhuo et al., 2012; Wang and Zheng, 2012). Also as an attractive method to address action recognition across camera views, the existing approaches

try to establish certain form of connection between the source and target views and use a commonly shared representation for both views when representing an action, then the trained classifier based on this shared representations can be adaptable for any instances represented in this shared form. For example, (Farhadi and Tabrizi, 2008; Farhadi et al., 2009) proposed to learn view-consistent split-based features from different views based on Maximum Margin Clustering. (Liu et al., 2011a; Li et al., 2012) advocated to construct a bilingual codebook as the shared representation from two view-dependent vocabularies using bipartite graph and (Zheng et al., 2012) tried to use transferable dictionary pairs to encourage the same action from different views to have similar sparse representations. Other emerging approaches include (Li and Zickler, 2012; Zhang et al., 2013) and (Huang et al., 2012). The former two exploited Grassmannian manifold and took the sampled points or the integral kernel between two views on the manifold as the shared representations, while the latter derived a correlated subspace where the shared representation for the same action from different views can be extracted. Although successful, these techniques still suffer from some drawbacks, e.g., extracting the view-consistent features is computational intensive, or the commonly shared representations are not accurate enough to guarantee the cross-view consistency. The Grassmannian manifold based approaches nicely characterized changes between source and target data, but did not explicitly exploit the statistical properties of the observed data.

2.2.3 Dictionary Learning Review

We used dictionary learning to investigate the statistical connection between different camera views, we thus also simply review the dictionary learning in this subsection. Dictionary learning, as a particular sparse model, is to find effective representation of data as a combination of a few typical patterns (atoms) learned from the data itself. The optimized dictionary is usually over-determined and composed by a relatively

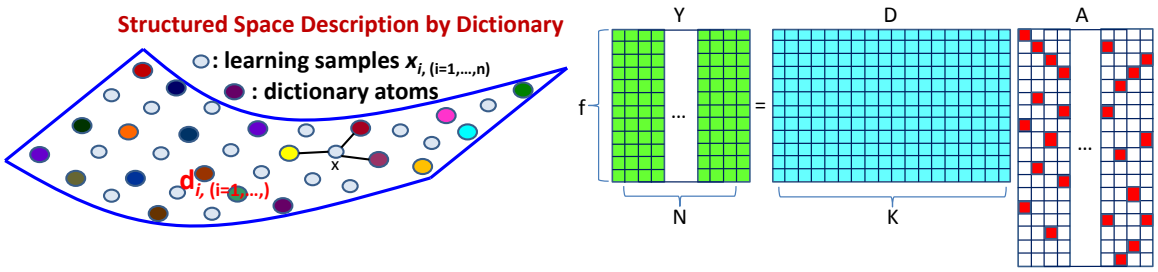

Figure 2.1: A graphic illustration of dictionary learning. Left, dictionary atoms are used for description of the data space structure; Right, feature samples are represented as sparse coefficient vectors.

large number of atoms, which effectively depict the structure of the data space similar to the landmarks on certain surface, as shown in the left of Figure 2.1.

Mathematically, given data samples $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_N] \in \mathbb{R}^{f \times N}$, find a dictionary $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_K]$ with columns number $K \ll N$, such that $\mathbf{y}_i$ is a sparse combination of a few columns in $\mathbf{D}$, as shown in the right of Figure 2.1.

$$\arg\min_{\mathbf{D},\mathbf{A}} = \sum_{n=1}^{N} ||\mathbf{y}_n - \mathbf{D}\mathbf{a}_n||^2 + \lambda|\mathbf{a}_n|_1, s.t., ||\mathbf{d}_k||^2 \leq 1, \forall k = 1, 2, ..., K \tag{2.1}$$

Eq. 2.1 indicates dictionary learning involves deriving the sparse codes $\mathbf{a}$, which is a high-dimensional vector with just a few elements being nonzero, and optimizing the dictionary $\mathbf{D}$, which always yields the sparse representations for the training data. An iterative optimization approach was proposed in (Lee et al., 2007; Mairal et al., 2009) for solving Eq. 2.1 with two main steps in each iteration. First, calculate the sparse coefficients in $\mathbf{A}$ by fixing dictionary $\mathbf{D}$. A bunch of methods were proposed, including OMP (Chen et al., 1998), l_1 minimization (Lee et al., 2007), shrinkage thresholding (Tibshirani, 1996; Beck and Teboulle, 2009), etc. Then, updating the dictionary $\mathbf{D}$ by fixing $\mathbf{A}$, it becomes a typical convex optimization problem can be solved by e.g., gradient decent (Aharon et al., 2006; Mairal et al., 2009).

In the past decade, the dictionary learning technique has been applied in a wide range of applications in computer vision area, such as image classification (Yang et al., 2010b; Jiang et al., 2013), face recognition (Wright et al., 2009; Li et al., 2013a),

saliency detection (Li et al., 2009b; Seo et al., 2014), objects visual tracking (Xue and Ling, 2009; Taalimi et al., 2015b), abnormal event detection (Cong et al., 2011; Tang et al., 2013), human action recognition (Luo et al., 2013b, 2014a; Wang et al., 2013d), biometric recognition (Taalimi et al., 2015a; Khorsandi et al., 2015), system monitoring (Wang et al., 2013b, 2014a), etc, as well as the cross domain recognition problems (Mehrotra et al., 2012; Qiu et al., 2012; Ni et al., 2013). For example, (Zhu and Shao, 2014) introduced a weakly-supervised cross domain dictionary learning approach that learns a reconstructive, discriminative and domain adaptive dictionary pair to bring the data from the original target domain and source domain into the same feature space.

2.3 Cross View Person Re-Identification

The purpose of person re-identification in a non-overlapping camera networks is to match pedestrian images observed in different camera views with visual features. It has important applications in video based surveillance, such as cross-camera tracking, multi-camera event detection, etc. The existing person re-identification approaches can be generally grouped into two categories: robust view-invariant feature extraction and supervised distance metric learning.

2.3.1 View Invariant Feature Design

The existing works on feature design and selection also can be further divided into unsupervised and supervised versions. Unsupervised approaches search for cross view invariant features via perceptual symmetry or certain prior assumptions (Cheng et al., 2011; Ma et al., 2012a; Liu et al., 2012a; Bazzani et al., 2012). For example, Farenzena et al (Farenzena et al., 2010) proposed the symmetry-driven accumulation of local features by exploiting the symmetry property. Zhao et al (Zhao et al., 2013b,a) used a salience model for patch matching such that the reliable and discriminative

matched patches can be picked for better performance. Liao et al (Liao et al., 2015) proposed to maximize the occurrence of each local pattern among all horizontal sub-windows to tackle viewpoint changes. Supervised approaches select the most effective features by certain criteria (Gray and Tao, 2008; Ma et al., 2012b). For example, Prosser et al (Prosser et al., 2010) formulated the person re-identification as a ranking problem, and extracted global feature weights based on an ensemble of RankSVM. Paisitkriangkrai et al (Paisitkriangkrai et al., 2015) improved the feature ensemble performance by learning the weights based on CMC curves. Recently, Wu et al (Wu et al., 2015) proposed an appearance model integrating camera viewpoint and human pose information novelly.

2.3.2 Distance Metric Learning

The basic idea behind metric learning is to learn a better similarity measure between the features from the same individual. In contrast, approaches that focus on metric learning usually extract features in a more straightforward way, e.g., color or texture histograms from predefined image regions. A lot of metric learning algorithms have been proposed recently (Zheng et al., 2011; Kostinger et al., 2012; Ma et al., 2014; Xiong et al., 2014; Li et al., 2015b). For example, Mahalanobis distance learning has been applied for the re-identification problem (Hirzer et al., 2012; Mignon and Jurie, 2012), as M-distance can implicitly model the transition in feature space between camera views. Besides, Pedagadi et al (Pedagadi et al., 2013) applied FDA together with PCA and LPP to derive a low dimensional yet discriminant subspace. Li et al (Li et al., 2013b) developed a locally-adaptive decision function (LADF) that jointly models a distance metric and a locally adaptive thresholding rule to achieved good performance. Dictionary learning techniques (Liu et al., 2014; Jing et al., 2015) are also proposed to bridge the appearance across two camera views with the assumption that the manifold of the local patches in spaces of two cameras are similar. Recently, Chen et al (Chen et al., 2015a) proposed an explicit polynomial kernel approach that

learns a similarity function to maximize the difference between the similarity score of the true and false image pairs.

Other than the two aforementioned main research branches, some other interesting and novel approaches also have been proposed for the person re-identification problem. For example, the deep learning framework was applied to exploit the information of the cross-input difference features by multiple layers of the neural network (Li et al., 2014b; Yi et al., 2014; Ahmed et al., 2015). The midd-level features, e.g., filters and semantic attributes were also investigated (Layne et al., 2012; Liu et al., 2012b). Zhao et al (Zhao et al., 2014) proposed to learn mid-level filters by mining the cross-view invariance in subsets of local patch features. Shi et al (Shi et al., 2015) proposed a novel approach for learning a semantic attribute model from existing fashion datasets, and adapting the resultant model to facilitate person re-identification.

2.3.3 Random Forest Review

We used random forest to learn adaptive local metrics for the person re-identification problem, we thus also simply review the random forest technique in this subsection. Random forest is an ensemble of decision trees for classification, regression and other tasks. It is operated by constructing a multitude of decision trees at training time and outputting as that is the mode of the classes (classification) or mean prediction (regression) of each individual tree (Trevor et al., 2008). The parameters vector Θ_k of each tree are learned using selection of different random subsets of training samples, random subset of features and a random split at each internal tree node. Although each decision tree sometimes prone to over-fitting to the training samples, however, with the combination of multiple trees in the forest with random learning, the over-fitting problem can be effectively relieved (Breiman, 2001).

Each decision tree in forest is a collection of internal nodes (split) and terminal nodes (leaf) in a hierarchical and binary structure, as shown in the left of Figure 2.2. Given a data sample as a feature vector $\mathbf{v} = (x_1, x_2, ..., x_d) \in \mathbb{R}^d$, the tree applies the

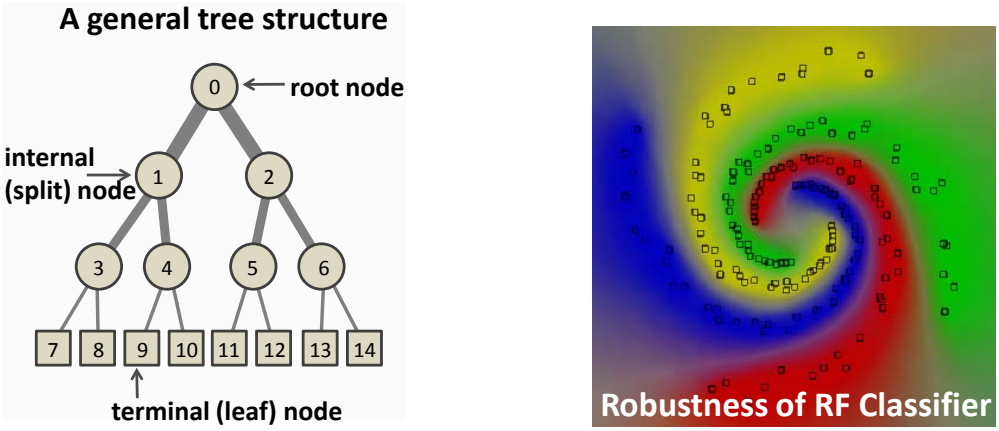

Figure 2.2: Left: structure of a decision tree. Right: an example case of non-linear classification based on random forest. The figures are from (Criminisi et al., 2011).

split function on the features x at each split node from the root recursively to the bottom leaf, where a predictor (e.g., classifier or regressor) generates a corresponding output for the input sample $\mathbf{v}$. In off-line training, given a set of training samples $\mathcal{S}_0 = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n\}$, the parameters Θ in the tree split functions are optimized to minimize a predefined energy or index function, such as purity or entropy, depending on the specific tasks (Criminisi et al., 2011). The trees in forest are randomly different one from another, leading to a de-correlation effect and strong generalization capability, therefore, random forest demonstrates strong prediction in both of classification and regression problems. An example case for non-linear classification problem is shown in the right of Figure 2.2. From the result, we can see the classifier is quite strong for this difficult classification on toy data.

Due to the strong generalization capability, random forest and its variations have been applied in a wide range of applications in computer vision area, such as image classification (Ristin et al., 2014, 2015), object detection (Schulter et al., 2013, 2014), object visual tracking (Tan and Ilic, 2014; Taixe et al., 2014), image denoise (Fanello et al., 2014), edge detection (Hallman and Fowlkes, 2015; Dollar and Zitnick, 2013), semantic segmentation (Bulo and Kontschieder, 2014; Shotton et al., 2008), human pose estimation (Dantone et al., 2013; Krupka et al., 2014). However, up to our best knowledge, random forest has not been investigated in cross domain recognition tasks.

2.4 Multi-Event Detection in Smart Grid

2.4.1 Smart Grid System

It has become essential that the wide-area situational awareness (WASA) systems can enable the monitoring of bulk power grid systems and provide critical information for understanding and responding to system disturbances and cascading blackouts. Event detection researches first began in 1980s on closely synchronized measurements that would allow direct measurement of the voltage phase angle at transmission level. As a result, Phasor Measurement Units (PMU) (Phadke and Thorp, 2008; Chow et al., 2009) have been gradually installed in substations that measure phasor at high voltage levels. As a member of the PMU family, the Frequency Disturbance Recorder (FDR) collects the instantaneous voltage phasor and frequency measurements at low-voltage distribution level using ordinary 120-V wall outlets. Based on these low-cost FDRs, a US-wide Frequency Monitoring Network (FNET) has thus been implemented (Zhong et al., 2005; Liu, 2006; Gardener and Liu, 2007). FNET now serves the entire North American power grid through advanced situational awareness techniques including, real-time event alert, accurate event spatial localization, animated event visualization, post event analysis and so on (Zhang et al., 2010).

2.4.2 Disturbance Event Analysis

There have been some research works reported so far conducting event analysis using real data collected from the FNET (Zhang et al., 2010; Li et al., 2010; Zhao et al., 2008; Xia et al., 2007; Gardner et al., 2006; Kook and Liu, 2011). In (Zhang et al., 2010; Li et al., 2010; Zhao et al., 2008), event detection was triggered if the rate of frequency change over a period of time exceeds an empirical threshold. Then, based on different event triggering times detected at multiple FDRs, event localization can be performed using approaches such as the geometrical triangulation (Xia et al., 2007) or the least-squares method (Gardner et al., 2006). (Kook and Liu, 2011) also

took advantage of the "order" of event detection from different FDRs and spatially localized events by finding the best matching between the actual detection order and that from simulations. Since the frequency signals collected from FDRs are usually corrupted by both white noise and impulsive noise, denoising becomes an important preprocessing step to guarantee the subsequent performance of event detection and localization. Representative works include adaptive median filter (Li et al., 2010), model fitting using adaptive Kalman filter (Zhao et al., 2008) or curve fitting (Wang et al., 2013c). Although successful, these state-of-the-art techniques can only handle disturbances caused by a single event. As far as we know, there are very few works reported for analysis of multi-event occurred in a cascading fashion, e.g., (Zhu and Giannakis, 2012) developed a algorithm for identifying multiple line outages by solving a sparse signal reconstruction problem via either greedy steps or coordinate descent iterations. The system disturbance reports from North American Electric Reliability Corporation (NERC) (NERC, 2010) have made it obvious that major disturbances typically involve a number of unlikely, unplanned events. Therefore, how to determine the number of events and identify the types of events with precise estimation of their occurring time becomes a very challenging problem.

We observe that when multiple events occur in smart grid system, the electromechanical waves generated will interfere with each other, and the measurement taken at a FDR sensor would more than likely be a "mixture" of multiple constituent event signals. Mixed measurements are frequently encountered in real-world applications, due to the resolution associated with discrete sampling and the effect of unknown sources, the measurements can rarely be pure. The existence of mixed measurements has brought the decomposition or unmixing technique to a wide array of applications. For example, in remote sensing area, due to the large footprint, a single pixel usually covers more than one type of ground constituent. Hence, the measured spectrum at a single pixel is a mixture of several ground cover spectra, where the pixel unmixing technique has been widely applied to sub-pixel object quantification (Wang and Qi, 2013; Guo et al., 2015), mineral elemental concentration estimation (Wang et al.,

2013a, 2014b), anomaly detection (Wang et al., 2015; Li et al., 2015a; Dong et al., 2009; Wang et al., 2009), etc. In other areas, including object detection (Luo and Qi, 2010), facial feature extraction (Guo and Qi, 2015), material modelling (Luo et al., 2012), system monitoring (Song et al., 2015), speech processing (Naqvi et al., 2012), biological microscopy (Hiraoka et al., 2002), etc. However, the signal of multi-event might not be a straightforward combination of some single-event signals because of the strong correlation among physical devices. Our work for multi-event signal analysis extracts some root patterns that are commonly shared in both of the single event domain and the multi-event domain, and compacts these root-patterns to a temporal awareable over-complete dictionary subtly. Then, the detection and recognition of the constituent events can be realized via non-negativity and sparsity constrained signal decomposition. Up to our best knowledge, this is the first realization for analysis of multi-event disturbances with high accuracy by making use of the information across different types of events.

Chapter 3

Human Action Recognition Across Camera Views

The same action may look quite different if viewed by different cameras from different angles. Therefore, it poses a great challenge for recognition tasks across different camera views. This chapter presents a novel approach to solving the problem of action recognition across camera views. Each action is represented based on a bag-of-visual-words model extracted from spatio-temporal features. Although the action representations are sensitive to view changes, our approach uses a reconstructable path to effectively bridge the high level semantic correspondences between actions in different views, such that the labelled training samples in any source view can be *translated* along the path into the target view for learning a view-adaptable classifier. In learning of the paths, a dictionary is assigned and optimized for each camera view to convert action representations into a sparsely represented space, and a linear mapping function is simultaneously optimized to bridge the gap between the source and the target spaces, such that each domain structure could be fully exploited and the discrimination among action categories can be well preserved after the translation. In addition, there might exist some samples cannot be directly used for the path learning - we also propose a scheme to investigate the hidden information embedded

in these seemingly useless samples, thus the stringency of the learning samples can also be relieved. The proposed approach is verified on the IXMAS action dataset under two working modes. The experimental results demonstrate that our approach achieves superior performance to the state-of-the-art with less strict requirements on the learning samples.

3.1 Introduction

Human action recognition is an essential task to many real world applications, such as visual surveillance, video retrieval, human-computer interaction, etc. Although the variability in human appearance, shape, posture and characteristic style in performing some motions makes the consistent description of a given action difficult, with the design of some discriminative features (Tran and Trivedi, 2008; Lin et al., 2009; Dollar et al., 2005; Luo and Qi, 2012; Luo et al., 2013b), many approaches have achieved very good recognition performance. However, the assumption in these works that all the actions captured for training and testing are from the same camera view is often violated because of the possible change of camera viewpoints. In practical scenarios, the same action may look quite different from a different angle, and hence difficult to be recognized because the magnitude of variations of action characteristics, which distinguishes one action from the others, may be even smaller than the variation caused by the change of viewpoints. Therefore, the recognition performance of the approaches only using these conventional action features tends to decrease dramatically.

Because of the lack of labelled training data, it is impractical to train individual classifiers for each camera view. We assume the labelled samples are only available in one or several *source* views while the testing actions in the *target* view might not be seen in advance. Under this assumption, most of the existing view-fixed action recognition approaches cannot be easily extended to recognize actions captured in a new *target* view. To handle this challenging cross-view action recognition problem,

there have been several kinds of approaches proposed, such as the view-invariant features (Juneio et al., 2008; Lewandowski et al., 2010), 3D model based features (Weinland et al., 2007; Li et al., 2007) and view inferable classifiers (Weinland et al., 2010; Wu and Jia, 2012). In addition, another emerging family of approaches is based on *transfer learning*, which encourages the actions recorded from different views to be represented by commonly shared representations (Farhadi and Tabrizi, 2008; Liu et al., 2011a) to achieve cross view invariance.

To enable the actions observed in target view to be directly recognized by a classifier trained by the set of labelled training actions merely from the source view, a straightforward idea is to translate those labelled training samples. Inspired by the photo-sketch synthesis research in (Wang et al., 2012), we propose an intuitive approach under transfer learning that it directly reconstructs the training samples from the source view into the target view for training a view-adaptable classifier. As illustrated in Figure 3.1, the same action from two different views are represented by their view-dependent vocabularies via the Bag of Visual Words (BoVW) model (Li and Perona, 2005). Then, along the indicated processing flow, which is referred to as the *reconstructable path*, any action representation from the source view is firstly converted to its sparse coding space, then mapped into the space of the target view, and finally reconstructed via the target dictionary as a *mirrored* instance in target view. In this way, if there are multi-source views available, the mirrored samples from different sources could be pooled together for learning a strong classifier in the target view. We refer to this cross view action recognition approach as the Reconstructable Path between individual View Dependent Representations (RP-VDR).

Under the framework of RP-VDR, two working modes are studied. The first one is *correspondence mode*. Similar to the work in (Liu et al., 2011a), where the unlabelled action samples are observed simultaneously in both of the source and target views, producing corresponding pairs, which can be used in learning of the reconstructable paths. Instead of requiring access to simultaneous multi-view observations of the same action instances, our approach can also work in the second mode, i.e., *partially labelled*

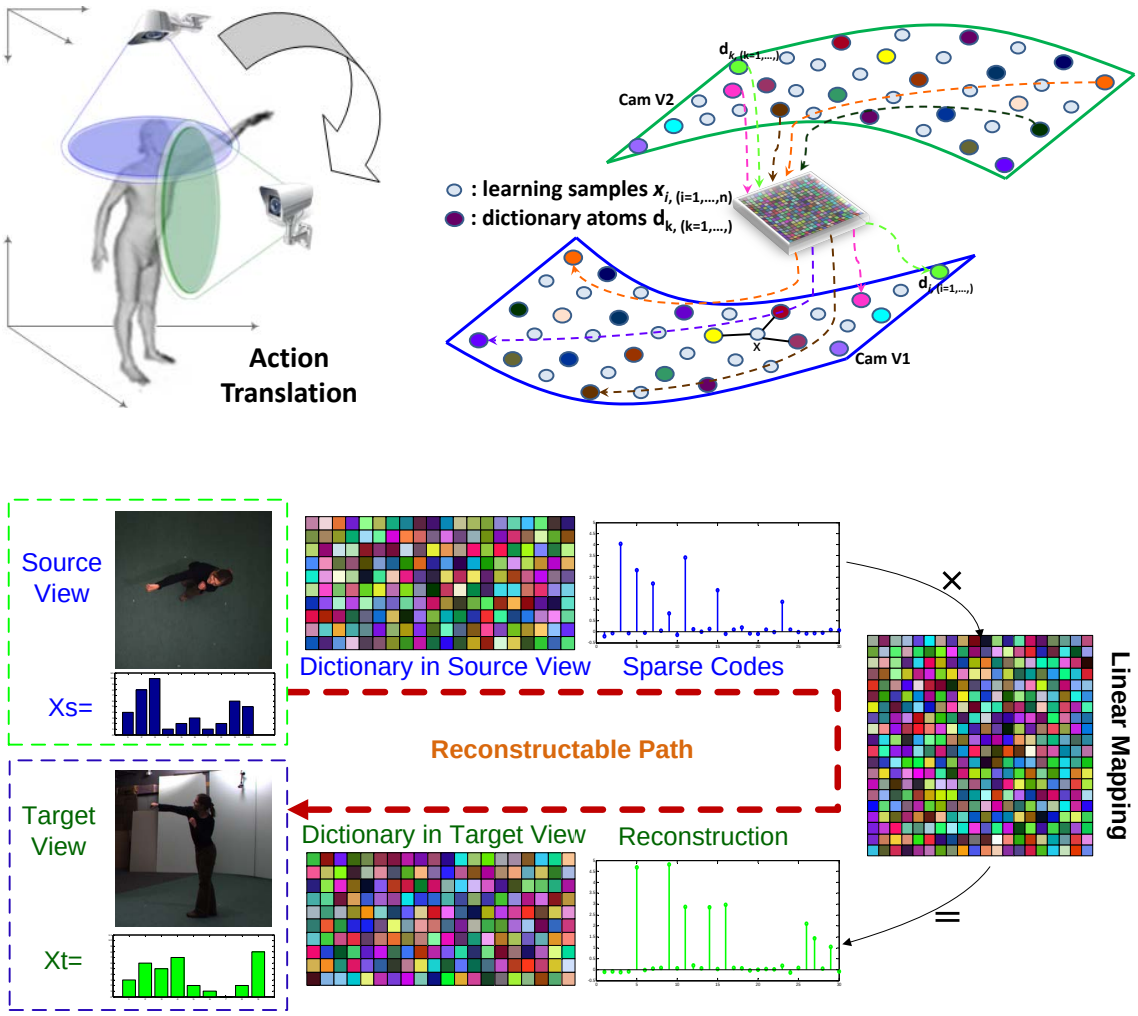

Figure 3.1: Top: illustration of how to correlate two camera view domains. Bottom: the process of one action reconstructed from the source view into the target view along a reconstructable path (marked as the dashed red line).

mode (Li and Zickler, 2012), that leverages weak supervision. Under this mode, the target view provides a few labelled samples while matched or corresponded instances within the source view are no longer exist.

In summary, the main contribution of this work is three-fold. First, the proposed RP-VDR directly reconstructs *(or mirrors)* action representations of the training samples from the source view into the target view, such that the statistical connection between the source and the target views can be exploited via the straightforward action-to-action correspondence. Second, the reconstructable path between any two views employs a dictionary for each view domain and an intermediate mapping function to bridge the semantic gap between them. In learning of the paths, our approach uses learning samples from both of the two different views to optimize the three terms alternately, such that the dictionaries are able to exploit the structure of each view domain thoroughly, resulting in better recognition for the mirrored actions with less discrimination. Third, RP-VDR also facilitates further exploitation of the hidden information embedded in the seemingly useless and isolated samples. For example, the unpaired samples in either the target or the source view under the first working mode, and the unlabelled samples in the target view under the second working mode. Usually, the number of paired instances or labelled samples in the target view is limited, and thus not sufficient for learning a strongly inferable reconstructable path. Therefore, how to make use of the hidden information is the key to improve the reconstruction capability in path learning, especially if only a small number of qualified learning samples are available. We also propose a companion approach, RP-VDRh, where the hidden information is also exploited to achieve comparable or even better recognition performance with much less strict path learning instances.

Methodologically, the most relevant works to our proposed work are (Zheng et al., 2012; Zheng and Jiang, 2013), which proposed to use transferable dictionary pairs to encourage the same actions from the *source* and the *target* views to have similar sparse representations (SpsRep). However, this approach still suffers from several limitations. First, the coupled learned dictionaries tend to over-fit the concatenated

action representations (ActRep) from two different views in learning phase, causing the two SpsReps derived from each single view separately in testing phase not being optimally close, although similar. In contrast, the mapping matrix between the view dependent dictionaries is able to well bridge this gap between the SpsReps. Second, although the SpsReps from different views of one action performed by the *same person* can be similar, due to the noise sensitive and unstable nature of the SpsRep, the same action performed by *different people* with small intra-class variations might result in quite large difference in the SpsReps. In contrast, we circumvent this disadvantage by reconstructing the actions from the source view into the target view directly, with guaranteed precision by the in-between mapping. Third, (Zheng et al., 2012; Zheng and Jiang, 2013) did not exploit the hidden information in the whole data set and demanded the availability of sufficient learning samples to assure good performance. In contrast, our approach can greatly reduce this stringent requirement on the path learning samples by making use of the seemingly useless and isolated samples to mine the hidden information.

3.2 Action Representation Reconstruction

Our work reconstructs view-dependent ActReps of the labelled training samples from the source view into the target view, and trains a classifier $\mathcal{C}$ based on these mirrored training samples for action recognition in the target view. As mentioned before, the ActRep reconstruction can be realized based on the optimized *reconstructable path*.

To obtain such a *path*, different settings are assumed in the two working modes. In *correspondence mode*, we are given pairwise action instances $(\mathbf{X}_s^p, \mathbf{X}_t^p) = \{\mathbf{x}_i^{ps}, \mathbf{x}_i^{pt}\}_{i=1:n}$, and also unpaired samples $\mathbf{X}_s^q = \{\mathbf{x}_i^{qs}\}_{i=1:n_s}$ in source view and $\mathbf{X}_t^q = \{\mathbf{x}_i^{qt}\}_{i=1:n_t}$ in target view. In the other *partially labeled mode*, the two sets of learning samples from the source and target views are not paired. Instead, besides the labelled samples in source view $\mathbf{X}_s^l = \{\mathbf{x}_i^{ls}\}_{i=1:n_s}$, we are also given a few labelled samples from target view $\mathbf{X}_t^l = \{\mathbf{x}_i^{lt}\}_{i=1:n_t}$, but $n_t \ll n_s$. To be clear, the n_s, n_t in the first mode denote the

number of unpaired samples while in the second mode denote the number of labelled samples in each view. The core of our approach is the learning of the *path* under both modes. We will first introduce how to optimize the path with paired instances ($\mathbf{X}_s^p,\ \mathbf{X}_t^p$) in the Sec. 3.2.2; and then introduce how to exploit the hidden information with unpaired $\mathbf{X}_s^q,\ \mathbf{X}_t^q$ in Sec. 3.2.3; finally, we introduce how to adapt the learning samples $\mathbf{X}_s^l,\ \mathbf{X}_t^l$ given in the other *partially labelled* mode to the same framework of path learning in Sec. 3.2.4.

3.2.1 Single View Dictionary Learning

Given a data matrix, $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \mathbb{R}^{m \times n}$, composed by n data points sampled in an m-dimensional feature space, the goal of dictionary learning is to learn the $\mathbf{D} = [\mathbf{d}_1, ..., \mathbf{d}_k] \in \mathbb{R}^{m \times k}$, and the corresponding sparse codes $\mathbf{A} = [\mathbf{a}_1, ..., \mathbf{a}_n] \in \mathbb{R}^{k \times n}$, thus the dataset $\mathbf{X}$ can be well approximated by $\mathbf{X} \approx \mathbf{DA}$. The problem can be described as minimizing the objective function $\mathcal{F}_n(\mathbf{D})$:

$$\min_{\mathbf{D},\mathbf{a} \in \mathbb{R}^{k \times n}} \frac{1}{n} \sum_{i=1}^{n} (\frac{1}{2} ||\mathbf{x}_i - \mathbf{D}\mathbf{a}_i||_2^2 + \lambda ||\mathbf{a}_i||_1 + \eta ||\mathbf{a}_i||_2^2) \tag{3.1}$$

where λ is a regularization parameter to trade off the sparsity of coefficients and the approximation of the input data, and η is a regularization parameter to encourage a group of correlated columns in $\mathbf{D}$ to have stable coefficients instead of just several ones. Though the objective function in Eq. (3.1) is not convex in both variables, it can be conveniently solved by alternately optimizing one variable while fix the other one via the algorithm in (Mairal et al., 2009).

In fact, dictionary learning for unsupervised clustering has also been exploited in (Ramirez et al., 2010; Sprechmann and Sapiro, 2010a), which showed that it can effectively investigate discrimination among categories by removing outliers and recover corrupted entries in feature vectors. In practical image or action recognition scenarios, the objects that are partially occluded or corrupted will not be easily

recognized (Wright et al., 2009). With an optimized dictionary $\mathbf{D}$, any action $\mathbf{x}_i$ can be represented as:

$$\mathbf{x}_i = \mathbf{D}\mathbf{a}_i + \mathbf{e}_i = \mathbf{x}'_i + \mathbf{e}_i \tag{3.2}$$

where the nonzero entries in $\mathbf{e}_i$ capture the corrupted part or noise information in $\mathbf{x}_i$. Thus, the $\mathbf{x}'_i$ can be treated as a *denoised representation* of $\mathbf{x}_i$ for better recognition performance, which will also be verified in Sec. 3.3.2 for evaluation of the single view action recognition performance and be taken as the baseline.

3.2.2 Learning the Reconstructable Path

As illustrated in Figure 3.1, the *reconstructable path* between any two camera views includes a dictionary for each view domain and a mapping function to relate the source and target views. There are two options to learn this path. *First*, we can learn each dictionary individually, such that $\mathbf{D}_s$ is derived by the set $\mathbf{X}_s = \{\mathbf{x}_i^s\}_{i=1:n_s}$, while $\mathbf{D}_t$ is derived by the set $\mathbf{X}_t = \{\mathbf{x}_i^t\}_{i=1:n_t}$, separately, where n_s, n_t denote the number of all the samples in source and target views, respectively. Then, the in-between linear mapping function $\mathbf{M}_{s2t}$ can be derived as:

$$\begin{aligned} &\mathbf{X}_s = \mathbf{D}_s\mathbf{A}_s + \mathbf{E}_s,\ \ \mathbf{X}_t = \mathbf{D}_t\mathbf{A}_t + \mathbf{E}_t \quad \rightarrow \\ &\mathbf{M}_{s2t} = \mathbf{A}_t \cdot \mathbf{A}_s^{+} = \mathbf{A}_t\mathbf{A}_s^T(\mathbf{A}_s\mathbf{A}_s^T + \epsilon\mathbf{I})^{-1} \end{aligned} \tag{3.3}$$

where $\mathbf{A}_s^{+}$ is the pseudo-inverse of $\mathbf{A}_s$, ϵ is a trivial value, and this learning mode is referred to as *separate learning*.

Second, the dictionaries $\mathbf{D}_s, \mathbf{D}_t$ and the linear mapping term $\mathbf{M}_{s2t}$ can be learned simultaneously in an alternate fashion. Therefore, the objective function is composed of three parts, the dictionaries $\mathbf{D}_s, \mathbf{D}_t$, the coefficients $\mathbf{A}_s, \mathbf{A}_t$ and the linear mapping function $\mathbf{M}_{s2t}$, respectively.

$$\min_{\mathbf{D}_s,\mathbf{D}_t,\mathbf{M}_{s2t},\mathbf{A}_s,\mathbf{A}_t} \{\mathcal{F}_n(\mathbf{D}_s) + \delta||\mathbf{A}_t - \mathbf{M}_{s2t}\mathbf{A}_s||_F^2 + \mathcal{F}_n(\mathbf{D}_t)\} \tag{3.4}$$

Then, Eq. (3.4) can be fully expressed as:

$$\begin{aligned}\min_{\mathbf{D}_{s,t},\mathbf{M}_{s2t},\mathbf{A}_{s,t}} \{\frac{1}{2}||\mathbf{X}_s - \mathbf{D}_s\mathbf{A}_s||_F^2 + \lambda_s||\mathbf{A}_s||_1 + \eta_s||\mathbf{A}_s||_F^2\} + \{\delta \\ ||\mathbf{A}_t - \mathbf{M}_{s2t}\mathbf{A}_s||_F^2\} + \{\frac{1}{2}||\mathbf{X}_t - \mathbf{D}_t\mathbf{A}_t||_F^2 + \lambda_t||\mathbf{A}_t||_1 + \eta_t||\mathbf{A}_t||_F^2\}\end{aligned} \tag{3.5}$$

where δ is the parameter to balance the linear mapping term. Although the objective function in Eq. (3.5) is not convex if optimize the three parts of variables jointly, each of them can be optimized alternately within a convex form by making the other two parts fixed.

The initial estimate of $\mathbf{D}_s$, $\mathbf{D}_t$ and $\mathbf{M}_{s2t}$ can be gained using the *separate learning*. With $\mathbf{D}_s$, $\mathbf{D}_t$ and $\mathbf{M}_{s2t}$ being fixed, $\mathbf{A}_s$ and $\mathbf{A}_t$ can be solved individually as follows. Notice that $\mathbf{M}_{t2s}$ is the inverse of $\mathbf{M}_{s2t}$.

$$\min_{\mathbf{A}_s} \frac{1}{2}||\mathbf{X}_s - \mathbf{D}_s\mathbf{A}_s||_F^2 + \delta||\mathbf{A}_t - \mathbf{M}_{s2t}\mathbf{A}_s||_F^2 + \lambda_s||\mathbf{A}_s||_1 + \eta_s||\mathbf{A}_s||_F^2 \tag{3.6}$$

$$\min_{\mathbf{A}_t} \frac{1}{2}||\mathbf{X}_t - \mathbf{D}_t\mathbf{A}_t||_F^2 + \delta||\mathbf{A}_s - \mathbf{M}_{t2s}\mathbf{A}_t||_F^2 + \lambda_t||\mathbf{A}_t||_1 + \eta_t||\mathbf{A}_t||_F^2 \tag{3.7}$$

Eqs. (3.6) (3.7) are both the multiple tasks sparse coding problems, thus $\mathbf{A}_s$, $\mathbf{A}_t$ can be both estimated efficiently and individually as the Lasso problem (Friedman et al., 2010; Mairal et al., 2009). Subsequently, $\mathbf{D}_s$ and $\mathbf{D}_t$ can be updated individually by fixing the other terms. Each of them forms a convex optimization problem:

$$\min_{\mathbf{D}_s} ||\mathbf{X}_s - \mathbf{D}_s\mathbf{A}_s||_2^2, \quad \min_{\mathbf{D}_t} ||\mathbf{X}_t - \mathbf{D}_t\mathbf{A}_t||_2^2 \tag{3.8}$$

The block-coordinated decent is used (Mairal et al., 2009) to update the $\mathbf{D}_s, \mathbf{D}_t$. For example, to solve $\mathbf{D}_s$ in Eq.(3.8), atoms in $\mathbf{D}_s$ are updated one by one. Let $\mathbf{d}_k$ be the k-th aton in $\mathbf{D}_s$. When updating atom $\mathbf{d}_k$, all the other atoms in $\mathbf{D}_s$ are fixed, and the first derivative of Eq.(3.8) over $\mathbf{d}_k$ can be derived as:

$$\begin{aligned}f(\mathbf{d}_k) = ||\mathbf{X}_s - \mathbf{D}_s\mathbf{A}_s||_F^2 = ||\mathbf{X}_s - (\mathbf{Q}_s + [0, \cdots, \mathbf{d}_k, \cdots, 0])\mathbf{A}_s||_F^2 \\ \rightarrow \nabla(f(\mathbf{d}_k)) = (-2\mathbf{X}_s + 2\mathbf{Q}_s\mathbf{A}_s + 2\mathbf{d}_k\alpha_{(k)})\alpha_{(k)}^\top\end{aligned} \tag{3.9}$$

where $\alpha_{(k)}$ is the k-th row in the matrix $\mathbf{A}_s$, and it is corresponding to the coefficients contributed by the atom $\mathbf{d}_k$. Matrix $\mathbf{Q}_s$ is of the same size as $\mathbf{D}_s$ and is the matrix after replacing the k-th column with zeros in $\mathbf{D}_s$. Therefore, the updated atom $\mathbf{d}_k$ can be calculated by setting $\nabla(f(\mathbf{d}_k))$ to zero, which is:

$$\begin{aligned} \mathbf{d}_k &= (\mathbf{X}_s - \mathbf{Q}_s\mathbf{A}_s)\alpha_{(k)}^\top / ||\alpha_{(k)}||_2^2 \\ &\rightarrow d_k = d_k / ||d_k||_2 \end{aligned} \tag{3.10}$$

Then, using the $\mathbf{A}_s, \mathbf{A}_t$ solved from Eqs. (3.6)(3.7), the linear mapping term $\mathbf{M}_{s2t}$ can be again updated by Eq. (3.3). Iterations are used to best optimize these variables. This learning method is referred to as *alternate learning*, and a pseudo-code of the *alternate learning* is provided as below.

Algorithm 1 Pseudo-code of Path Learning via RP-VDR

Input: $(\mathbf{X}_s^p, \mathbf{X}_t^p) = \{\mathbf{x}_i^{ps}, \mathbf{x}_i^{pt}\}_{i=1:n}$: paired action samples from two camera views.
K: the number of iterations in optimization.

Output: $\mathbf{D}_s$: space description of source view. $\mathbf{D}_t$: space description of target view.
$\mathbf{M}_{s2t}$: inter-between linear mapping term.

1: **Initialization**: $\mathbf{D}_s, \mathbf{D}_t, \mathbf{M}_{s2t}$ based on Eq. (3.3)
2: **while** $k \leq$ K **do**
3: Optimize $\mathbf{A}_s, \mathbf{A}_t$ via Eqs. (3.6)(3.7), which can be transformed as:
$\min_{\mathbf{A}_s} \frac{1}{2} \left\| \begin{bmatrix} \mathbf{X}_s \\ \sqrt{\delta}\mathbf{A}_t \end{bmatrix} - \begin{bmatrix} \mathbf{D}_s \\ \sqrt{\delta}\mathbf{M}_{s2t} \end{bmatrix} \mathbf{A}_s \right\|_F^2 + \lambda_s ||\mathbf{A}_s||_1 + \eta_s ||\mathbf{A}_s||_F^2$
4: Optimize $\mathbf{D}_s, \mathbf{D}_t$ via Eq. (3.8). Atoms in $\mathbf{D}_s$ (or $\mathbf{D}_t$) are updated one by one. Let $\mathbf{d}_k$ be the k-th atom in $\mathbf{D}_s$. When updating atom $\mathbf{d}_k$, all the other atoms in $\mathbf{D}_s$ are fixed, and the first derivative over $\mathbf{d}_k$ can be derived as: $||\mathbf{X}_s - \mathbf{D}_s\mathbf{A}_s||_F^2 = ||\mathbf{X}_s - (\mathbf{Q}_s + [\mathbf{0}, ..., \mathbf{d}_k, \mathbf{0}, ...])\mathbf{A}_s||_F^2 \rightarrow \nabla(f(\mathbf{d}_k)) = (-2\mathbf{X}_s + 2\mathbf{Q}_s\mathbf{A}_s + 2\mathbf{d}_k\alpha_{(k)})\alpha_{(k)}^\top$ where $\alpha_{(k)}$ is the k-th row in the matrix $\mathbf{A}_s$, corresponding to the coefficients contributed by the atom $\mathbf{d}_k$.
Matrix $\mathbf{Q}_s$ is of the same size as $\mathbf{D}_s$ and is the matrix after replacing the k-th column with zeros in D_s. Therefore, the updated atom $\mathbf{d}_k$ can be calculated by setting $\nabla(f(\mathbf{d}_k))$ to 0. As, $\mathbf{d}_k = (\mathbf{X}_s - \mathbf{Q}_s\mathbf{A}_s)\alpha_{(k)}^\top / ||\alpha_{(k)}||_2^2$, $\mathbf{d}_k = \mathbf{d}_k / ||\mathbf{d}_k||_2$.
5: **end while**
6: Optimize $\mathbf{M}_{s2t}$ term via Eq. (3.3).
7: **End for**

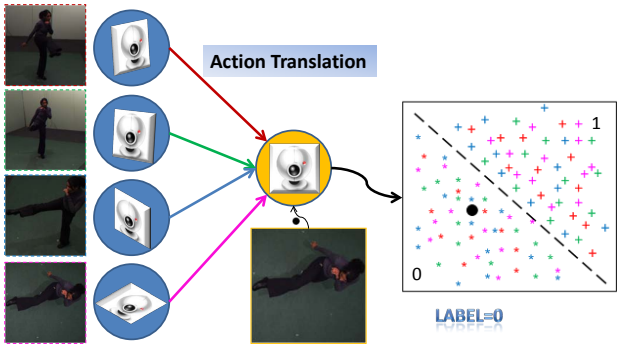

Figure 3.2: Multiple sources mixed-training of an action classifier in one target view.

Based on a reconstructable path, the dictionaries and the linear mapping function work together to correlate the source and target views efficiently. For a given labelled action sample $\mathbf{x}_s$ from the source view, it can be reconstructed as:

$$\mathbf{x}_s = \mathbf{D}_s\mathbf{a}_s + \mathbf{e},\ \bar{\mathbf{a}}_t = \mathbf{M}_{s2t}(\mathbf{a}_s),\quad \rightarrow \bar{\mathbf{x}}_t = \mathbf{D}_t\bar{\mathbf{a}}_t \tag{3.11}$$

Here, the reconstruction $\bar{\mathbf{x}}_t$ is sufficient if the path is learned using *separate learning*. However, it puts more weight on $\mathbf{D}_s$ if the path is learned using the *alternate learning*, since only the fidelity of $\mathbf{x}_s = \mathbf{D}_s\mathbf{a}_s + \mathbf{e}$ is considered. To balance the reconstruction in the latter learning mode, we can take the preliminary reconstructed $\bar{\mathbf{x}}_t$ as a prototype, which will be paired with $\mathbf{x}_s$ into Eqs. (3.6)(3.7) again for an alternate optimization for $\mathbf{a}_s$ and $\overrightarrow{\mathbf{a}_t}$. Then, the further optimized $\overrightarrow{\mathbf{a}_t}$ can be used to produce a more precise mirror of $\mathbf{x}_s$ as $\overrightarrow{\mathbf{x}_t} = \mathbf{D}_t\overrightarrow{\mathbf{a}_t}$. If *multiple*-source camera views are available to recognize unknown actions from one target camera view, as shown in Figure 3.2, we can simply reconstruct these labelled training samples from each source view into the one common target view, and all the reconstructed training samples will be used together to train a unified classifier.

3.2.3 Exploitation of Hidden Information

Learning of the reconstructable paths usually requires sufficient number of paired samples to gain enough capacity for ActRep inference. However, in real applications

we may be only given a small number of paired samples for the path learning while the other samples existing in different views are not recorded simultaneously. As shown in Sec. 3.3, the performance of cross view action recognition will degrade if the reconstructable paths are learned with only a small number of paired samples. To make the path learning be less contingent on the number of paired samples, we propose a companion approach, RP-VDRh, to exploit the hidden information embedded in the seemingly useless and isolated samples based on the prior that any action sample should belong to a specific action category and any two of the samples belong to the same action category from two different views could be matched as a pair to assist the path learning.

Given paired learning samples $(\mathbf{X}_s^p, \mathbf{X}_t^p) = \{\mathbf{x}_i^{ps}, \mathbf{x}_i^{pt}\}_{i=1:n}$, unpaired (or called as isolated) samples $\mathbf{X}_s^q = \{\mathbf{x}_i^{qs}\}_{i=1:n_s}$ in source view, and $\mathbf{X}_t^q = \{\mathbf{x}_i^{qt}\}_{i=1:n_t}$ in target view, we expect to learn the paths via the objective function in Eq. (3.12). Compared to the objective function in Eq. (3.5), the additional term here is $\mathbf{W} \in \mathbb{R}^{n_t \times n_s}$, which is used to manipulate the isolated samples in set $\mathbf{X}_t^q$ to be matched with the isolated samples in set $\mathbf{X}_s^q$. Each column in $\mathbf{W}$ is a vector with only one element being non-zero.

$$
\begin{aligned}
&\min_{\mathbf{D}_s,\mathbf{D}_t,\mathbf{M}_{s2t},\mathbf{A}_s,\mathbf{A}_t} \{\delta||\mathbf{A}_t^p - \mathbf{M}_{s2t}\mathbf{A}_s^p||_F^2\} + \{\frac{1}{2}||\mathbf{X}_s^p - \mathbf{D}_s\mathbf{A}_s^p||_F^2 + \lambda_s||\mathbf{A}_s^p||_1 + \eta_s||\mathbf{A}_s^p||_F^2\} \\
&+ \{\frac{1}{2}||\mathbf{X}_t^p - \mathbf{D}_t\mathbf{A}_t^p||_F^2 + \lambda_t||\mathbf{A}_t^p||_1 + \eta_t||\mathbf{A}_t^p||_F^2\} \\
&+ \{\delta||\mathbf{A}_t^q - \mathbf{M}_{s2t}\mathbf{A}_s^q||_F^2\} + \{\frac{1}{2}||\mathbf{X}_s^q - \mathbf{D}_s\mathbf{A}_s^q||_F^2 + \lambda_s||\mathbf{A}_s^q||_1 + \eta_s||\mathbf{A}_s^q||_F^2\} \\
&+ \{\frac{1}{2}||\mathbf{X}_t^q\mathbf{W} - \mathbf{D}_t\mathbf{A}_t^q||_F^2 + \lambda_t||\mathbf{A}_t^q||_1 + \eta_t||\mathbf{A}_t^q||_F^2\}
\end{aligned}
\tag{3.12}
$$

RP-VDRh first performs the path learning based on Eq. (3.5) by using the paired samples in $(\mathbf{X}_s^p, \mathbf{X}_t^p)$ only. With the preliminary learnt $\mathbf{D}_s$, $\mathbf{D}_t$, the unpaired samples from both the source and the target views can be transformed into the sparse code domain. Then, among these unpaired and isolated samples: $\mathbf{X}_s^q$, $\mathbf{X}_t^q$, the similarity between any pair of them can be measured, and the ones that either performed by the same person or with very small intra-class variation will be selected. Therefore,

in the first *correspondence* mode, these isolated and unpaired samples in $\mathbf{X}_s^q$ and $\mathbf{X}_t^q$ can be matched into pairs via $\mathbf{W}$. A transitionary matrix $\overline{\mathbf{W}} \in \mathbb{R}^{n_t \times n_s}$ is calculated with the value of each element being the pairwise similarity measure:

$$\begin{aligned} \overline{\mathbf{W}}_{i,j} &= \frac{\mathbf{a}_i \times (\mathbf{M}_{s2t} \cdot \mathbf{a}_j)}{||\mathbf{a}_i||\ ||(\mathbf{M}_{s2t} \cdot \mathbf{a}_j)||}, i \in [1, n_t], j \in [1, n_s] \\ \mathbf{W}_{i,j} &= \begin{cases} 1 & \text{if} \quad \overline{\mathbf{w}}_{i,j} < \overline{\mathbf{w}}_{k \neq i,j} \\ 0 & \text{else} \end{cases} \quad s.t. \sum_i \mathbf{W}_{i,j} = 1 \end{aligned} \tag{3.13}$$

where $\mathbf{a}_i$, $\mathbf{a}_j$ are the sparse codes transformed from $\mathbf{x}_i^{qt}$, $\mathbf{x}_j^{qs}$. Then, the dictionaries $\mathbf{D}_s, \mathbf{D}_t$ can be further optimized via Eq. (3.12) with additional knowledge from the matched samples through the matching matrix $\mathbf{W}$. Similarly, Eq. (3.12) can be solved as Eq. (3.5) once these seemingly useless samples $\mathbf{X}_s^q$ and $\mathbf{X}_t^q$ from two views are matched into pairs. In addition, this pair matching process will be repeated several times to explore the information from these isolated samples thoroughly, thus the learned dictionaries can be refined gradually. A pseudo-code of the RP-VDRh is also provided as below.

Algorithm 2 Pseudo-code of Path Learning via RP-VDRh

Input: $(\mathbf{X}_s^p, \mathbf{X}_t^p) = \{\mathbf{x}_i^{ps}, \mathbf{x}_i^{pt}\}_{i=1:n}$: paired action learning samples.
$\mathbf{X}_s^q = \{\mathbf{x}_i^{qs}\}_{i=1:n_s}$, $\mathbf{X}_t^q = \{\mathbf{x}_i^{qt}\}_{i=1:n_t}$: isolated action samples.
K: the number of iterations.

Output: $\mathbf{D}_s, \mathbf{D}_t, \mathbf{M}_{s2t}$: descriptions of the views and the intermediate term.

1: **Initialization**:
Preliminary learning of $\mathbf{D}_s, \mathbf{D}_t, \mathbf{M}_{s2t}$ based on $(\mathbf{X}_s^p, \mathbf{X}_t^p)$ only via pseudo-code (1)
2: **while** $k \leq$ K **do**
3: Solve $\mathbf{A}_s^q, \mathbf{A}_t^q$ via the preliminarily learned dictionaries $\mathbf{D}_s, \mathbf{D}_t$.
4: Pairwise matching of samples in set $\mathbf{X}_s^q$, $\mathbf{X}_t^q$ via $\mathbf{W}$ and Eq. (3.13).
5: Optimize terms $\mathbf{D}_s, \mathbf{D}_t, \mathbf{M}_{s2t}$ via Eq. (3.12).
6: **end while**

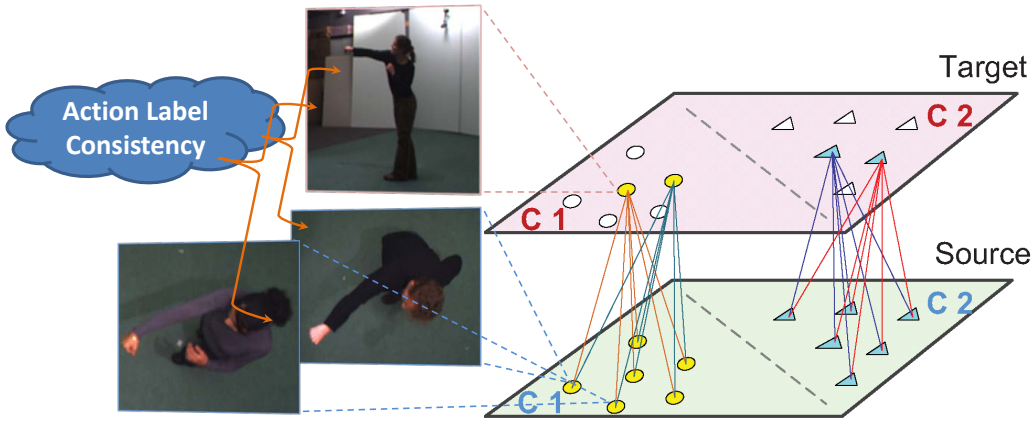

Figure 3.3: Pairwise combination process via action label-consistency between different camera views, the white icons are unlabelled and colored icons are labelled.

3.2.4 Using Partially Labelled Target Samples

Under the *partially labelled* working mode, we are given two sets of labelled learning samples $\mathbf{X}_s^l = \{\mathbf{x}_i^{ls}\}_{i=1:n_s}$ and $\mathbf{X}_t^l = \{\mathbf{x}_i^{lt}\}_{i=1:n_t}$, from the source and the target views, respectively. However, they are not in pair, even any personal-correspondence is not exist, meaning the persons appeared in source view will not appear in target view. Only a few ($n_t \ll n_s$) samples ($\mathbf{X}_t^l$) from target view are provided with labels, which is the key to exploring the cross view connection. We propose an effective approach that adopts a pairwise combination process. Given the labelled samples from target view, we can pair each sample in the source view with every possible sample in the target view of the same action category, just as shown in Figure 3.3. This idea is feasible since ideally any action representation of the same category should be identical. We thus can pair any two actions of the same category from two different views without personal correspondence, but just the action label's consistency. This prior enables the seemingly coarse pairwise combination process to produce a quite comprehensive set of paired instances for learning of the reconstructable paths via RP-VDR. In addition, to exploit the information embedded in the other unlabelled samples $\mathbf{X}_t^u$ in target view for better path learning, we match these unlabelled target samples with some of the source samples via RP-VDRh. Similarly, in learning of the classifier, these

labelled training samples will be reconstructed from source view into target view via the reconstructable path, such that they can also be pooled together with the given few labelled target samples to learn a stronger classifier for better action recognition.

3.3 Experiments

3.3.1 Experimental Setup and Rules

Dataset: We test our approach based on the popular IXMAS multiple views action dataset (Weinland et al., 2007), which contains 11 categories of daily actions. Each action is performed 3 times by 12 actors taken from 5 *different views* including 4 side views and 1 top view. Therefore, there are totally 396 action videos under each camera view. From the example actions, "kick" and "punch", shown in Figure 3.4, we can find two challenges that: *first*, under the same camera view, e.g., the 1st column, the same action performed by different actors has certain intra-class variations, as the different actor may perform the same action with some different orientations; *second*, across the different camera views, e.g., the 1st row, the same action performed by the same actor looks quite different from different viewpoints.

Learning Parameters and Classification Setup: In learning of the reconstructable paths, the parameter δ is set as 2, and the $\lambda_t, \lambda_s, \eta_t, \eta_s$ are all set as 0.01 and the number of atoms in D_s, D_t are set as 300. These parameters are insensitive and defined empirically, therefore, a little bit change of these values would not affect the performance much. In the phase of action recognition across cameras, we use a 6-fold cross-validation for the evaluation, thus six multi-class SVM classifiers are trained based on the labelled samples from the source view and utilized to recognize actions from target view. The 6-fold cross-validation also guarantees that the actors whose actions will be tested in the target view are all excluded in the training of the SVM classifier, so as to avoid any unfair recognition. The SVM with the histogram intersection kernel (Maji et al., 2008) is used as our classifier.

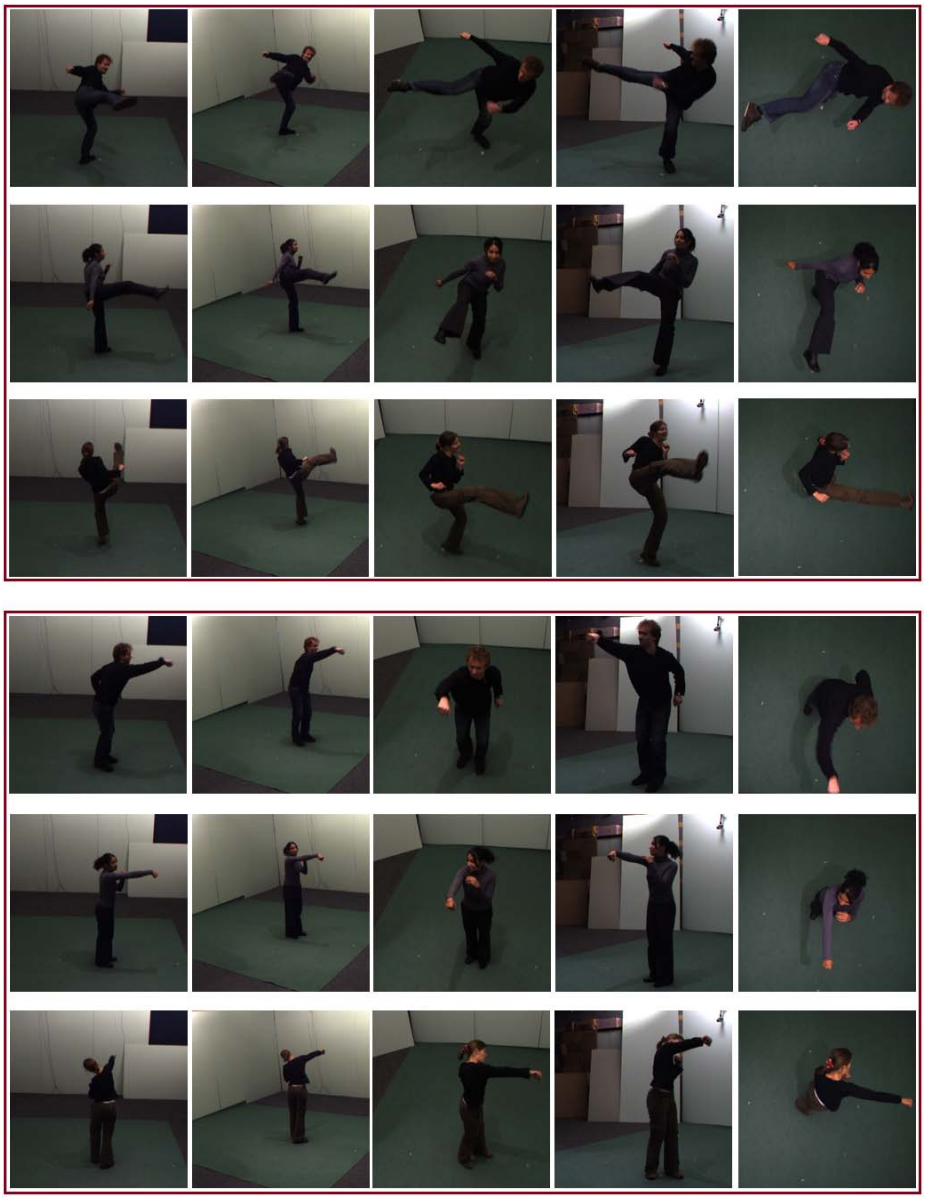

Figure 3.4: Two example actions 'kick' and 'punch' taken from five (each row) different camera viewpoints (0∼4), performed by 3 (each column) different actors.

Action Features for Representation: For fair comparison purpose, we adopt the same action descriptors as used in the literature (Liu et al., 2011a; Wu and Jia, 2012; Li and Zickler, 2012; Huang et al., 2012; Zheng et al., 2012; Zhang et al., 2013). Specifically, the actions are represented by a concatenation of a spatio-temporal local interest points feature (Dollar et al., 2005) and a global shape-flow feature (Tran and Sorokin, 2008), based on the bag-of-visual-words model (Li and Perona, 2005). The two types of features serve as complementary descriptions to characterize actions.

The local motion features of each action are extracted from some interest points, which are detected as the local maximal response by a 2D Gaussian filter followed by a 1D-Gabor filter. We use the same parameter settings as in (Dollar et al., 2005) for the two filters as $\sigma = 2$ and $\tau = 1.5$, respectively, and at most 200 interest points will be extracted from each action video. Then, the spatio-temporal volumes around these points are extracted and described by the gradient-based descriptor. PCA is applied to reduce the dimension of the volume descriptors to be 100. The volume descriptors from all training actions are quantized into 1000 visual-words by K-means clustering and thus each action is represented as a histogram of 1000 visual-words.

The global shape-flow features of each action are extracted in each frame. Three-channel features are extracted including horizontal optical-flow, vertical optical-flow and silhouette (Tran and Sorokin, 2008). PCA is also applied to reduce the dimension of each feature. The temporal information is also taken into account by concatenating the features from neighboring frames into description of the current frame. Similarly, the feature descriptors from all training action videos are quantized into 500 visual-words, and then each action is represented as a histogram of the 500 visual-words. Finally, by concatenating the local and global features, we obtain the final form of action representation as a histogram of 1500 dimension.

Design of Experiments: We ***first*** evaluate the denoised action representation DnBoVW, as illustrated in Sec. 3.2.1, in the same view. Since our approach mirrors the labeled action samples from source view into target view for training a classifier, the recognition performance by the classifier trained by samples from the same view

can be taken as the Baseline, and our objective is to make the recognition accuracies across camera views be as close to that under the same view as possible. ***Second***, we evaluate the cross view action recognition under the *correspondence mode*, including optimization by *separate learning* and *alternate learning*, the robustness of RP-VDRh if the number of the pairwise learning instances decreases, as well as comparison to the other existing works. ***Third***, we evaluate the action recognition performance under the *partially labelled mode*, including the robustness to the different proportions of available labelled target samples, and also the comparison to the other existing works. ***Fourth***, we evaluate the recognition performance if multi-source views are available under the aforementioned two working modes.

Furthermore, we use the *leave-one-action-class-out* scheme (Liu et al., 2011a; Li and Zickler, 2012; Zhang et al., 2013) in evaluation. The scheme means that we only consider one action category in test of each round, where we call the left-out action as the "*orphan action*". All the videos in the "*orphan action*" category are excluded in learning of the view-dependent visual words and also the reconstructable paths, such that this scheme is able to test the generalization capability of the dictionaries for a new unseen action.

3.3.2 Single View Action Recognition

Since the recognition of actions in the same view of training can be taken as a baseline for the performance of cross view recognition, we firstly look into the effectiveness of the *denoised* BoVW action representation (DnBoVW) as introduced in Eq. (3.2). We take 300 (5/6) *non-orphan* actions to learn the dictionary under each view individually in each round, and the averaged recognition accuracies of tests on each *orphan action* are shown in the diagonal entries of Table 3.1. Compared to the original BoVW, we can find that the DnBoVW performs better since it is able to recover the corrupted motion information from the action samples.

Table 3.1: Diagonal entries: action recognition in the same view, top: BoVW, bottom: DnBoVW. Non-diagonal entries: cross view action recognition via BoVW, reconstructable Paths with SL and AL on IXMAS. (Row: source; Column: target)

B./% S./A.	Cam0	Cam1	Cam2	Cam3	Cam4
Cam0	88.5, 93.4,	8.03, 42.7, 95.5	9.09, 38.4, 90.4	12.9, 45.5, 87.1	12.4, 48.7, 90.2
Cam1	10.2, 44.4, 93.0	85.2, 96.8,	10.3, 47.8, 89.4	11.8, 43.2, 82.8	10.9, 53.5, 89.4
Cam2	9.76, 44.7, 87.9	10.6, 54.8, 90.4	90.8, 93.9,	8.48, 47.5, 88.6	9.15, 55.6, 91.2
Cam3	10.1, 44.7, 87.6	8.48, 44.4, 83.9	10.0, 36.2, 92.9	88.7, 94.8,	9.09, 48.3, 83.6
Cam4	9.76, 59.6, 81.6	8.10, 60.4, 87.9	11.3, 60.4, 89.4	11.8, 52.8, 77.5	85.3, 92.7,
Ave.	9.94, 48.4, **87.8**	8.52, 50.6, **89.4**	10.1, 45.6, **90.5**	11.3, 47.2, **84.1**	10.4, 51.5, **88.6**

3.3.3 Pairwise Cross View Recognition

Correspondence Mode: We then look into the performance of action recognition across pairwise camera views in the correspondence mode. We still follow the popular data separation scheme *leave-one-action-class-out* for a fair comparison. Similarly, all videos of the "*orphan action*" are excluded in learning.

We start with the performance comparison between using the BoVW representation, the results of RP-VDR learned by *separate learning* (SL) and *alternate learning* (AL). 300 (5/6) *non-orphan* actions are used in each round for path learning, and the averaged testing results are shown in Table 3.1. It is not surprising that the performance of BoVW is much worse than the other two since there is no connection established between the source and target views. In learning of the reconstructable paths, since the AL considers both the fidelity of dictionary terms and the fidelity of linear mapping term simultaneously in each iteration, it can balance the relationship between the three terms and perform much better than the SL, which puts much less weight on the fidelity of the linear mapping term than on the other two. Although the recognition accuracy across cameras via the reconstructable paths is still a bit lower

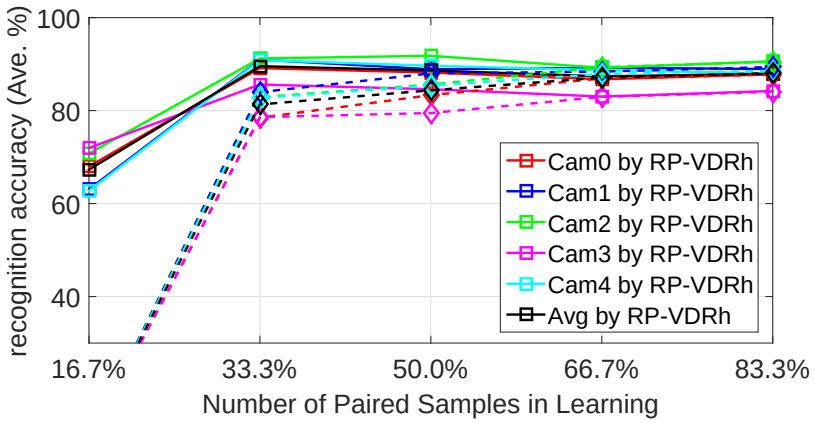

Figure 3.5: Comparison between RP-VDR and RP-VDRh: averaged recognition accuracies cross pairwise views with different number of paired instances when each view is taken as a target view.

compared to that under the same view using DnBoVW, the result is quite promising that it already achieves or exceeds the level of using the BoVW only for recognition under the same single view.

We then vary the number of paired action instances in learning of the paths to investigate the effect of changing this parameter. The number of randomly selected action instances in path learning decreases sequentially from 300 (5/6 of the *non-orphan* samples) to 240(4/6), 180(3/6), 120(2/6), 60(1/6). The averaged recognition performance across pairwise views for different target views are shown in Figure 3.5, from which we can observe that the accuracies drop accordingly if the number of paired instances decreases. In contrast, the accuracies can be well kept at the similar level when RP-VDRh is utilized in path learning for mining the hidden information, demonstrating the robustness of RP-VDRh to the number of paired learning instances. We also observe that, with RP-VDRh, using only about 30% paired learning instances performs even better than using more paired instances. This might attribute to the matching process in RP-VDRh permitting two samples performed by different actors to be matched, leading to even stronger generalization for the reconstructable path. However, if the number of paired instances decreases further, the recognition accuracy

Table 3.2: Performance of different approaches for cross view action recognition on IXMAS dataset with paired instances (correspondence mode). The accuracy values in each tuple are from approaches in (Farhadi et al., 2009), (Zheng et al., 2012), (Liu et al., 2011a), (Li and Zickler, 2012), (Zhang et al., 2013) and ours, respectively.

%	Cam0	Cam1	Cam2	Cam3	Cam4
Cam0		79.0, 94.3, 79.9 81.8, 86.3, **94.5**	79.0, 61.5, 76.8 88.1, 93.1, 91.4	68.0, 83.8, 76.8 87.5, 91.5, 87.9	76.0, 69.8, 74.8 81.4, 85.4, **89.7**
Cam1	72.0, 92.9, 81.2 87.5, 90.5, **93.4**		74.0, 71.5, 75.8 82.0, 87.8, **89.2**	70.0, 69.2, 78.0 92.3, 91.3, 82.6	66.0, 59.4, 70.4 74.2, 83.4, **91.7**
Cam2	71.0, 67.1, 79.6 85.3, 90.4, 89.2	82.0, 79.2, 76.6 82.6, 84.4, **92.7**		76.0, 88.1, 79.8 82.6, 87.1, **90.4**	72.0, 68.9, 72.8 76.5, 81.6, **91.4**
Cam3	75.0, 82.7, 73.0 82.1, 86.3, **91.4**	75.0, 63.8, 74.1 81.5, 85.2, **87.9**	79.0, 78.3, 74.4 80.2, 85.3, **92.9**		76.0, 58.9, 71.2 70.0, 77.2, **90.7**
Cam4	80.0, 74.6, 82.0 78.8, 85.9, 82.6	73.0, 53.8, 68.3 73.8, 76.2, **88.9**	73.0, 80.3, 74.0 77.7, 84.5, **91.4**	79.0, 66.7, 71.1 78.7, 83.1, 81.6	
Ave.	74.5, 79.3, 79.0 83.4, 88.3, **89.2**	77.3, 72.8, 74.7 79.9, 83.0, **91.0**	76.3, 72.9, 75.2 82.0, 87.7, **91.2**	73.2, 77.0, 76.4 85.3, 88.3, 85.6	72.5, 64.3, 71.2 75.5, 81.9, **90.8**

will drop dramatically, since the paths will not be informative enough for inference with so few instances in the preliminary learning.

Finally, our approach is compared to other existing works. For fair comparison, all the actions are represented by similar local/global spatio-temporal features, and all of the recognition tests are under the 6-fold cross-validation manner with SVM classifiers. The cross-validation guarantees that any instance in test will not be used in the classifier's training. With about 30% paired instances for the path learning, the averaged recognition accuracies for each pairwise cross source-target views on each of the orphan actions are shown in Table 3.2. We can observe that our approach achieves improvement over the state-of-the-art on most of the cross view pairs (14/20), except when camera 3 is involved as a target view. Meanwhile, it is most desirable that our approach also achieves very good recognition accuracies when camera 4 is involved as either source or target view. As shown in Figure 3.4, camera 4 captures totally different action appearance from the top viewpoint, so its recognition accuracy is more meaningful for evaluating an approach for cross view recognition. Benefited from the optimization scheme for learning of the reconstructable paths, the structure

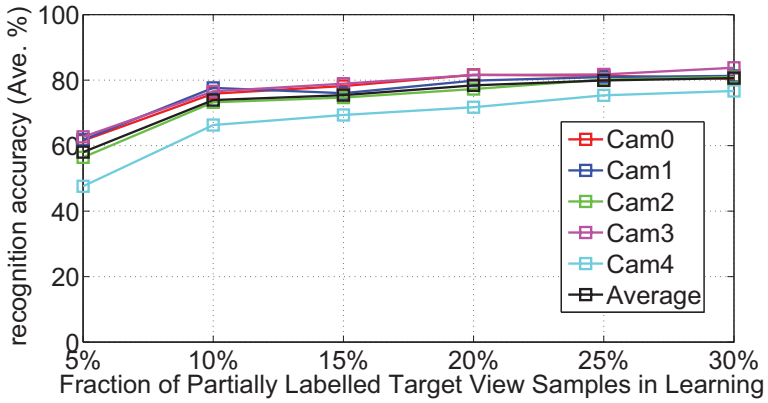

Figure 3.6: Averaged recognition accuracies across pairwise views when different proportion of samples are labeled in target view. Each line represents the averaged accuracies if one view is taken as the target view.

information in both the source view and the target view can be thoroughly exploited, such that the *discrimination* among actions can be well preserved after the action representation reconstruction. Also note that our performance is based on the most basic SVM with a histogram intersection kernel instead of the more advanced MKL-SVM (Li and Zickler, 2012).

Partially Labelled Mode: We also follow (Li and Zickler, 2012; Zhang et al., 2013) to consider a semi-supervised scenario, where a small proportion of labelled action samples in target view will also be provided. Note that all the corresponding samples of the labelled target samples in the source view are all excluded to strictly enforce that there are no instances in correspondence. In learning of the paths, we only consider the *alternate learning*, which provides superior action representation inference. After all the labelled actions from the source view be reconstructed into the target view, 6 SVM classifiers under the 6-fold cross validation are used for recognition performance evaluation.

To study the robustness to the number of partially labelled samples given from target view, we also vary the fraction of the randomly selected labelled target samples in increment of 5% to 30%. The averaged recognition accuracies using RP-VDPh across pairwise views for each target view are shown in Figure 3.6. From the results,

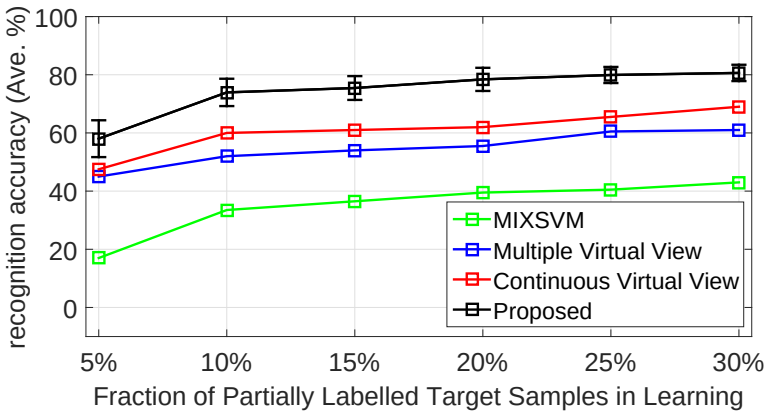

Figure 3.7: Averaged recognition accuracies across pairwise views with different proportion of labelled samples from target view. Comparisons are with the MIXSVM (Bergamo and Torres, 2010), Virtual View(Li and Zickler, 2012; Zhang et al., 2013).

we can observe that the performance of our approach under partially labelled mode is consistently high even with a small proportion of labelled target samples (10%).

The performance of our approach is also compared to other existing works under the same recognition framework. For all the approaches, up to 30% action samples in target view are randomly selected and labelled for learning. As shown in Figure 3.7, the results indicate that our approach achieves quite notable improvement. We also show quantitative comparison in Table 3.3. The experimental results show that our approach significantly outperforms the existing works in all the pairwise cross view tests when 30% labelled action samples are provided from the target view, with about 11.6% improvement to the state-of-the-art performance. Even with just 10% labelled action samples in the target view, the performance of our approach is still constantly superior to the others with 30% labelled target view samples in most of the pairwise cross view tests (16/20). Once again, the results clearly show the effectiveness of our approach on the partially labelled mode.

3.3.4 Multi-Source View Recognition

The benefits gained from multi-source views are investigated in this subsection. Each camera view is selected as the target view and the rest 4 serve as multi-source views.

Table 3.3: Performance comparison for cross view action recognition on IXMAS dataset if only a few labeled actions available in target view with no correspondence (partially labeled mode). The accuracies in each tuple are from (Bergamo and Torres, 2010), (Li and Zickler, 2012), (Zhang et al., 2013) with 30% labeled samples and our proposed with 5%, 10% and 30% labeled samples, respectively.

%	Cam0	Cam1	Cam2	Cam3	Cam4
Cam0		36.8, 63.6, 71.5 60.9 *77.0*, **81.8**	46.8, 60.6, 68.9 51.5 *71.5*, **81.6**	42.7, 61.2, 67.3 59.6 *73.5*, **84.9**	36.7, 52.6, 64.2 45.2 *64.2*, **75.0**
Cam1	39.4, 61.0, 70.5 58.3 **80.1**, *79.6*		51.8, 62.1, 69.8 57.1 *77.0*, **80.0**	45.8, 65.1, 74.2 63.4 *77.8*, **81.8**	40.2, 54.2, 62.3 45.7 *67.4*, **76.8**
Cam2	49.1, 63.2, 67.8 65.7 *73.5*, **81.1**	49.4, 62.4, 71.8 60.6 *78.3*, **80.3**		45.0, 71.7, *79.2* 62.1 76.8, **84.1**	46.9, 58.2, 66.5 51.3 *70.0*, **74.5**
Cam3	39.3, 64.2, 68.7 60.1 *75.3*, **83.1**	42.5, 71.0, *80.0* 60.4 78.0, **81.6**	51.2, 64.3, 70.4 58.8 *72.5*, **81.6**		38.9, 56.6, *63.8* 48.2 63.6, **80.6**
Cam4	40.3, 50.0, 55.4 62.1 *74.8*, **78.0**	42.5, 59.7, 67.3 65.4 *77.3*, **81.6**	40.4, 60.7, *72.6* 58.3 72.0, **81.1**	40.7, 61.1, 68.0 65.7 *78.3*, **84.3**	
Ave.	42.6, 59.6, 65.6 61.6 *75.9*, **80.4**	42.8, 64.2, 72.7 61.8 *77.7*, **81.3**	47.5, 61.9, 70.4 56.4 *73.2*, **80.9**	43.5, 64.8, 72.2 62.7 *76.6*, **83.8**	40.7, 55.4, 64.2 47.6 *66.3*, **76.7**

The SVM classifiers are trained based on the reconstructed training samples from all the four source views into the common target view, and the SVM classifiers are still trained within the 6-fold cross-validation. The averaged accuracies of our approach in *correspondence mode* and *partially labeled mode* are compared to other existing works in Table 3.4 and Table 3.5. If compare Tables 3.4 and 3.5 to Tables 3.2 and 3.3, we can observe that a notable accuracy improvement can be gained by fusing multiple source views under both modes. If compare our results to the other works under the *correspondence mode*, our approach achieves about 8% accuracy improvement on average. Especially under the *partially labelled mode*, our approach performs better ($\approx$ 5% improvement) even with just 10% labelled action samples given in the target view, while the other existing works used 30% labelled target samples. We also study the effect caused by the number of paired instances or labeled target samples in these two modes. As shown in Figure 3.8, we can find that our approach is very robust to the number of action samples involved in paths learning for the two working modes.

We further plot the recognition accuracy of each action category in the aforementioned two working modes in Figure 3.9. From results of the correspondence mode,

Table 3.4: Recognition accuracy with multiple source views in correspondence mode, at least 30% learning action pairs are used in the other existing works in comparison.

%	Cam0	Cam1	Cam2	Cam3	Cam4	Ave.
Ours (30.0%)	**96.0**	**96.5**	**96.0**	**92.9**	**94.7**	**95.2**
Avg. PairWise	89.2	91.0	91.2	85.6	90.8	89.6
Ours (16.7%)	88.9	82.3	93.2	94.0	93.2	90.3
(Zhang et al., 2013) 2013	89.2	85.6	88.0	90.7	83.6	87.4
(Wu and Jia, 2012) 2012	92.4	95.4	93.2	87.1	62.9	86.2
(Liu et al., 2011a) 2011	86.6	81.1	80.1	83.6	82.8	82.8
(Li and Zickler, 2012) 2012	85.1	82.1	82.2	85.7	77.6	82.6

Table 3.5: Recognition accuracy with multi-source views in partially labeled mode given 10%, 20%, 30% labeled samples in target view, while the other works used 30%.

%	Cam0	Cam1	Cam2	Cam3	Cam4	Ave.
Ours 30%	86.4	84.9	83.6	85.4	78.8	83.8
Ours 20%	83.3	83.1	81.1	85.1	77.3	81.8
Avg. PairWise	80.4	81.3	80.9	83.8	76.7	80.6
Ours 10%	**78.5**	**80.6**	**72.0**	**77.0**	**68.2**	**75.3**
(Zhang et al., 2013) 2013	66.4	73.5	71.0	75.4	66.4	70.5
(Li and Zickler, 2012) 2012	62.0	65.5	64.5	69.5	57.9	63.9
(Bergamo and Torres, 2010) 2010	46.4	44.2	52.3	47.7	44.7	47.1

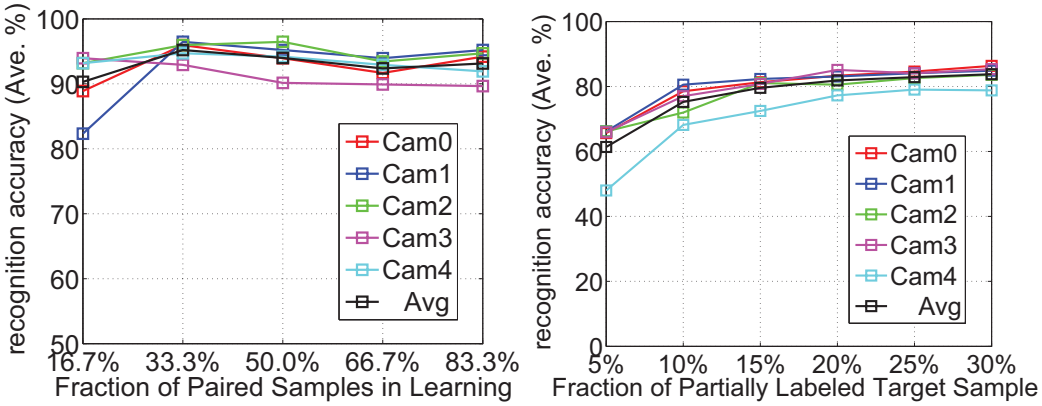

Figure 3.8: Averaged cross view action recognition accuracies with different number of samples used in learning when multiple source views available. Left, *corresponding mode*; Right, *partially labeled mode*

we can observe that most of the actions under each target view are classified correctly, except for the action 'walk' when 'Cam3' serves as the target view. This is mainly due to the mis-recognition of the action 'walk' as the action 'turn around' under this view because these two actions share very similar motion pattern. If we look into the action videos, we can find the action 'walk' is just 'turn a big round'. From the results of the partially labelled mode, we however observe that the recognition accuracies of all the *arm* related actions are not satisfactory, mainly due to the challenges caused by the similar motion patterns in these actions. In addition, it is also worth looking into the confusion matrices when camera 4 serves as the target view, which generally performs worse than the other cameras as the target view in previous works. The results in Figure 3.10 indicate that most of the actions are recognized with high accuracy except for the action 'wave' in the correspondence mode, which are partially mis-recognized as 'check watch'. This may be because that camera 4 records actions from the top viewpoint, where actions involving the arms, e.g., 'wave' and 'check watch', tend to have quite similar features. Again, the *arm* related actions are also tend to be mis-recognized in the partially labelled mode from the top camera view, e.g., 'scratch', 'wave' and 'check watch'.

3.3.5 Orientation Recognition and Processing Speed

Given a testing action sample $\mathbf{x}$ in real applications, we may not know which view it can best adapt to. With the dictionaries learned for each camera view, we may further estimate the action's orientation, then take the most suitable view as the target view for action recognition. In this way, we may realize a kind of relaxed version of view-invariant action recognition. Suppose we have J views, we can define a new matrix $\mathcal{D} = [\mathbf{D}_1, ..., \mathbf{D}_J]$ as the concatenation of the J dictionaries from J views. Then, the $\mathbf{x}$ can be represented as: $\mathbf{x} = \mathcal{D}\psi + e$, where $\psi = [\mathbf{a}_1^{\mathrm{T}}, ..., \mathbf{a}_J^{\mathrm{T}}]^{\mathrm{T}}$. Using only the coefficients associated with the i-th view, we can approximate the given action $\mathbf{x}$ by $\mathbf{D}_j \cdot \mathbf{a}_j$. We

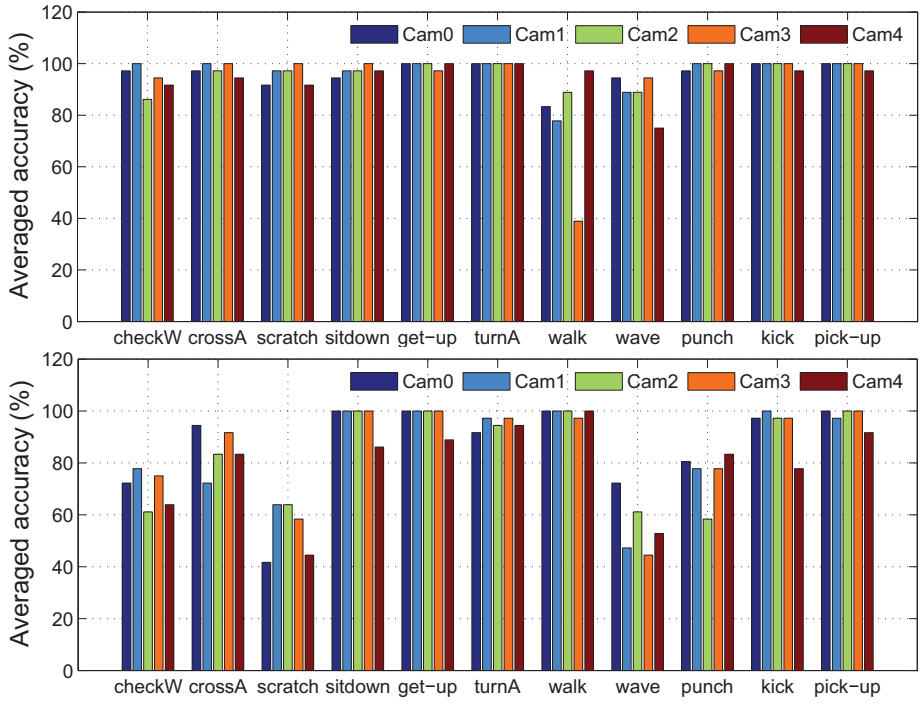

Figure 3.9: Recognition accuracy on each action category for target view if multi-source views available. Top: *correspondence mode*; Bottom: *partially labeled mode*.

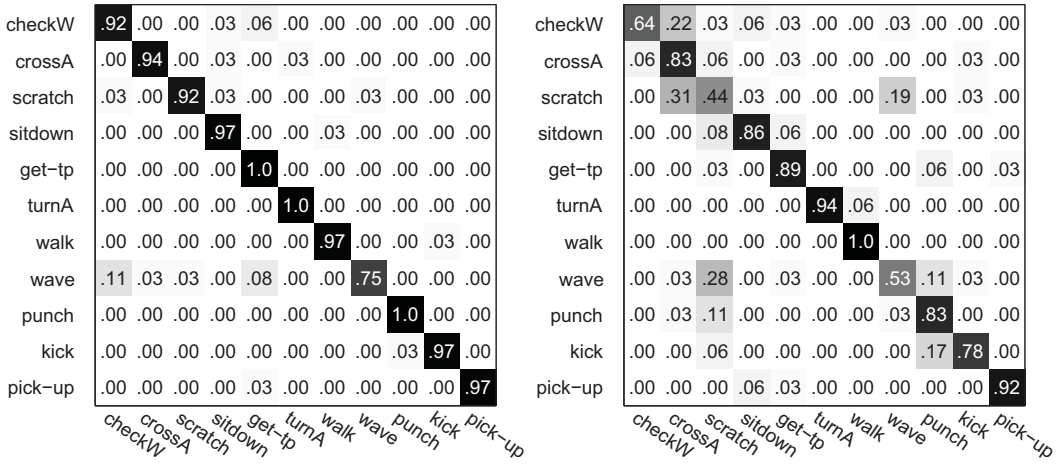

Figure 3.10: Confusion matrices if camera 4 (from top viewpoint) serves the target view under *correspondence mode* (left) and *partially labeled mode* (right), respectively.

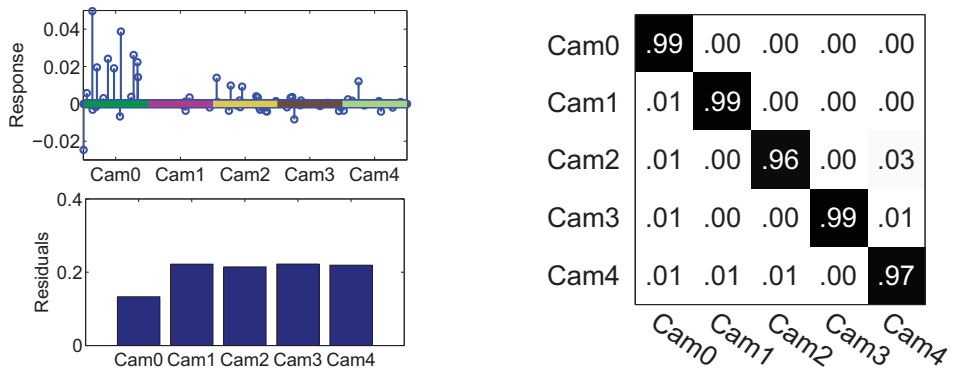

Figure 3.11: Actions orientation recognition. Left: coefficients (top) and residual (bottom) of an example action from camera 0; Right: confusion matrix of actions orientation recognition from 5 camera views.

then estimate the orientation of $\mathbf{x}$ based on these approximations by assigning it to the best adaptable view that minimizes the residual $r_j(\mathbf{x}) = ||\mathbf{x} - \mathbf{D}_j \cdot \mathbf{a}_j||_2$.

For example, we compute the coefficients and residuals of an action from camera view 0 as shown in the left of Figure 3.11, from which we can see that the coefficients mainly appear in the range of camera 0 and the minimal residual also appears in camera 0. For evaluation, we perform orientation recognition for all the 396×5 actions from five camera views under the *leave-one-action-class-out* scheme, the confusion matrix of recognition is shown in the right of Figure 3.11, from which we can find that the recognition of action orientation is quite accurate.

We also measure the processing speed for recognizing an action in target view on an ordinary laptop with I7 core, 4G memory in Matlab environment. It is true that the path learning will cost a bit long time, about 27.79 seconds for learning one path. However, both the path learning and the classifier training are all set up before test, we can recognize an action very fast in the test phase. The main portion of time cost in test phase is the action's feature extraction and representation, which cost about 0.832 second, while recognizing an action in a target view with trained classifier only costs 0.014 and 0.018 second with 1 or 4 source views, respectively. Thus, the total time consumption of our approach to recognize an action in a target view is about 0.85 second.

3.4 Summary

This work presented a novel approach for action recognition across camera views. Different from previous works searching for view-independent or commonly shared representations across camera views, our approach effectively exploited high-level semantic (label) correspondence among actions via action representation reconstruction. With a dictionary assigned to each camera view to fully exploit domain structural information, and a simultaneously optimized linear mapping function for bridging the semantic gap between the camera views, a reconstructable path was established between any two camera views. Through the path, the labelled action samples from any source view can be reconstructed (mirrored) into target view for training a strong SVM classifier in the target domain. In addition, the proposed RP-VDRh approach is able to exploit the hidden information from the unpaired samples or unlabelled samples, therefore the stringency of the path learning samples can be greatly relieved. Extensive experiments on the multiple view IXMAS dataset confirmed the use of our approach for improved performance of cross view action recognition upon the state-of-the-art.

Chapter 4

Person Re-Identification Across Camera Views

In this chapter, we propose a new approach for the person re-identification problem, discovering the correct matches for a query pedestrian image from a set of gallery images. It is well motivated by our observation that the overall complex inter-camera transformation, caused by the change of camera viewpoints, person poses and view illuminations, can be effectively modelled by a combination of many simple local transforms, which guides us to learn a set of more specific local metrics other than a fixed metric working on the feature vector of a whole image. Given training images in pair, we first align the local patches using spatially constrained dense matching. Then, we use a decision tree structure to partition the space of the aligned local patch-pairs into different configurations according to the similarity of the local cross-view transforms. Finally, a local metric kernel is learned for each configuration at the tree leaf nodes in a linear regression manner. The pairwise distance between a query image and a gallery image is summarized based on all the pairwise distance of local patches measured by different local metric kernels. Multiple decision trees form the proposed random kernel forest, which always discriminatively assign the optimal local metric kernel to the local image patches in re-identification. Experimental results over the

Figure 4.1: Samples of pedestrian images observed in different camera views in person re-identification. Each pedestrian has a different pose variation in the four examples between two cameras.

public benchmarks demonstrate the effectiveness of our approach for achieving very competitive performances with a relatively simple learning scheme.

4.1 Introduction

Person re-identification is to recognize the same person across a network of cameras with non-overlapping views. It is important for video surveillance by saving a lot of human effort on exhaustively searching for a person from large amounts of video sequences, e.g., the large scale pedestrian retrieval Loy and Tang (2009) and the wide scale multi-camera tracking Wang (2013). However, this is also a fairly challenging problem since the appearance of the same person may vary greatly in different camera views, due to the significant variations in camera viewpoints, illuminations, person poses, occlusions and backgrounds, etc. In addition, a surveillance camera usually observes hundreds of people in one day, many of which have similar appearances, therefore generating a lot of false alarms for the query image. See Figure 4.1 for some typical difficult examples.

Since the camera views are significantly disjoint making the temporal transition between cameras very large, the appearance is exploited solely in most existing works. In literature, two lines of approaches have been proposed to tackle this problem. The *first* line concentrated on the development of viewpoint quasi-invariant local features, e.g., color Gray and Tao (2008), texture Xiong et al. (2014) or gradient Zhao et al. (2013b), as well as robust feature ensembles. However, these feature based methods still suffer from illumination changes, human shape deformations and difficulty of multi-feature ensembles. The *second* line is to learn a parametric distance metric to enforce features from the same individual to be closer than that from different individuals Zheng et al. (2013); Li et al. (2013b); Pedagadi et al. (2013); Xiong et al. (2014), also known as the metric learning (ML). However, ML usually deals with feature vectors of a complete image in learning of the metric. Although effective, this distance metric may not be the optimal to work well on certain local parts of each person image.

In the re-identification problem, image regions typically undergo both geometric transformation due to camera viewpoint changes and photometric transformation due to illumination variations. However, different regions suffer differently to these two transformations, e.g., the smooth pure color regions suffer less while the texture or high gradient patches suffer more. In addition, the pose changes from one camera to another one vary for different people, there is no fixed pattern, i.e., $45^\circ \rightarrow 90^\circ$ or $0^\circ \rightarrow 45^\circ$, to describe the diverse pose changes, shown as in Figure 4.1. Therefore, the configuration of person images are multi-modal even if the people are observed in the same camera view. To fully formulate the overall inter-camera transformation $\mathcal{F}$, it must be a sophisticated non-linear function with a large number of unknown parameters. Obviously, single transform or uni-modal metric function might not be the optimal to tackle the problem. Thus some of the recent works used kernel tricks to do ML in a non-linear kernel space Xiong et al. (2014); Chen et al. (2015a) or adopted nested formulations as in deep learning framework Li et al. (2014b); Ahmed et al. (2015), which is usually time-consuming in model training.

Our work is mainly motivated by the above observations. Suppose a specific local metric can be learned from a small group of local image patch-pairs that share a consistent cross-view transform, not only the metric learning task becomes much easier, the combination of these specific metrics is also more effective to further ensure the pairwise distances of images from the same individual can be better minimized. Comparing to the deep learning architecture Li et al. (2014b); Ahmed et al. (2015), which approximates the overall transformation $\mathcal{F}$ as a series of nested functions with the distance metric defined as $\mathcal{D}(\mathbf{x},\mathbf{y}) = d_K\left(...d_2\left(d_1(\mathbf{x},\mathbf{y})\right)\right)$, where $\mathbf{x},\mathbf{y}$ are the representations of two images from two different camera views, respectively, we try to partition out all the local transforms and decompose the overall transformation $\mathcal{F}$ into many independent sub-functions f_k, then our new distance metric is defined as $\mathcal{D}(\mathbf{x},\mathbf{y}) = \sum_{x,y}\{d_1(x,y) + d_2(x,y) + ... + d_K(x,y)\}$, where x, y are features of local patches from the image pair $\mathbf{x},\mathbf{y}$, respectively, i.e., some segments of the feature vectors of $\mathbf{x},\mathbf{y}$. However, each d_k *only* works on a specific kind of the local patches from each image.

The main purpose of this work is to learn specific metric kernels for different local image patches in measure of the pairwise distance. We propose a novel random kernel forest (RKF) based on the *consistent patch-to-patch transform* criteria for person re-identification. Our *main contribution* is the use of a highly efficient decision forest that is trained to discriminatively predict which kernel should be applied to measure the pairwise distance of any two given image patches. As shown in Figure 4.2, the tree structure jointly partitions the space of local patch-pairs from all the training image pairs into a set of sub-spaces at each tree leaf, where the transform of the local patches between cameras is simplified and consistent. *Furthermore*, a simple linear kernel can be learned at each leaf to describe the specific transform $f_{k,k=1,...,\mathrm{K}}$, such that the distance between any true patch-pair will be minimized in d_k. Combining with multiple decision trees in the forest, the model also effectively avoids over-fitting during training. *Finally*, since the decision tree recursively and jointly partitions the patch-pair space solely based on the thresholds on features, it is very fast in

learning and prediction. Extensive experimental results demonstrate the effectiveness of our approach for achieving very competitive performance while adopting a relatively simple learning scheme.

4.2 Method

Random forests Criminisi and Shotton (2013) is a well-known tree based classifier ensemble. It has been widely used in many computer vision problems. In our work, the random forest has been strategically designed to **decompose** the multi-modal inter-camera transformation into simple and independent uni-modal transforms.

4.2.1 Transformation Model

Traditional machine learning problems try to learn a category specific probability distribution or a decision boundary to answer which category a given sample belongs to. In contrast, the person re-identification problem deals with image pairs and tries to determine whether a pair of samples are from the same category or not. Formally, for a pair of image samples represented by $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, respectively, each of which corresponds to a class label $\mathcal{C}(\mathbf{x})$ and $\mathcal{C}(\mathbf{y})$, we need to decide whether they are from the same category, i.e., $\mathcal{C}(\mathbf{x}) = \mathcal{C}(\mathbf{y})$, or not. The ability of dealing with unseen categories is the key for person re-identification, since most of the testing samples are from unseen persons which do not exist in the training set. The proposed approach still follows the distance metric learning framework. Given a set of N training pedestrian image pairs $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N\}$, which are observed by two disjoint camera $\mathcal{X}$ (cam$\mathcal{X}$) and camera $\mathcal{Y}$ (cam$\mathcal{Y}$), our goal is to learn a distance metric $\mathcal{D}(\mathbf{x}_i, \mathbf{y}_i)$ that any pair of two samples from the same person generates the smallest distance.

Mathematically, the gist of metric learning is to learn a projection P and find a common subspace to measure the pairwise distance, e.g., $||\mathbf{Px} - \mathbf{Py}||^2 = (\mathbf{x} -$

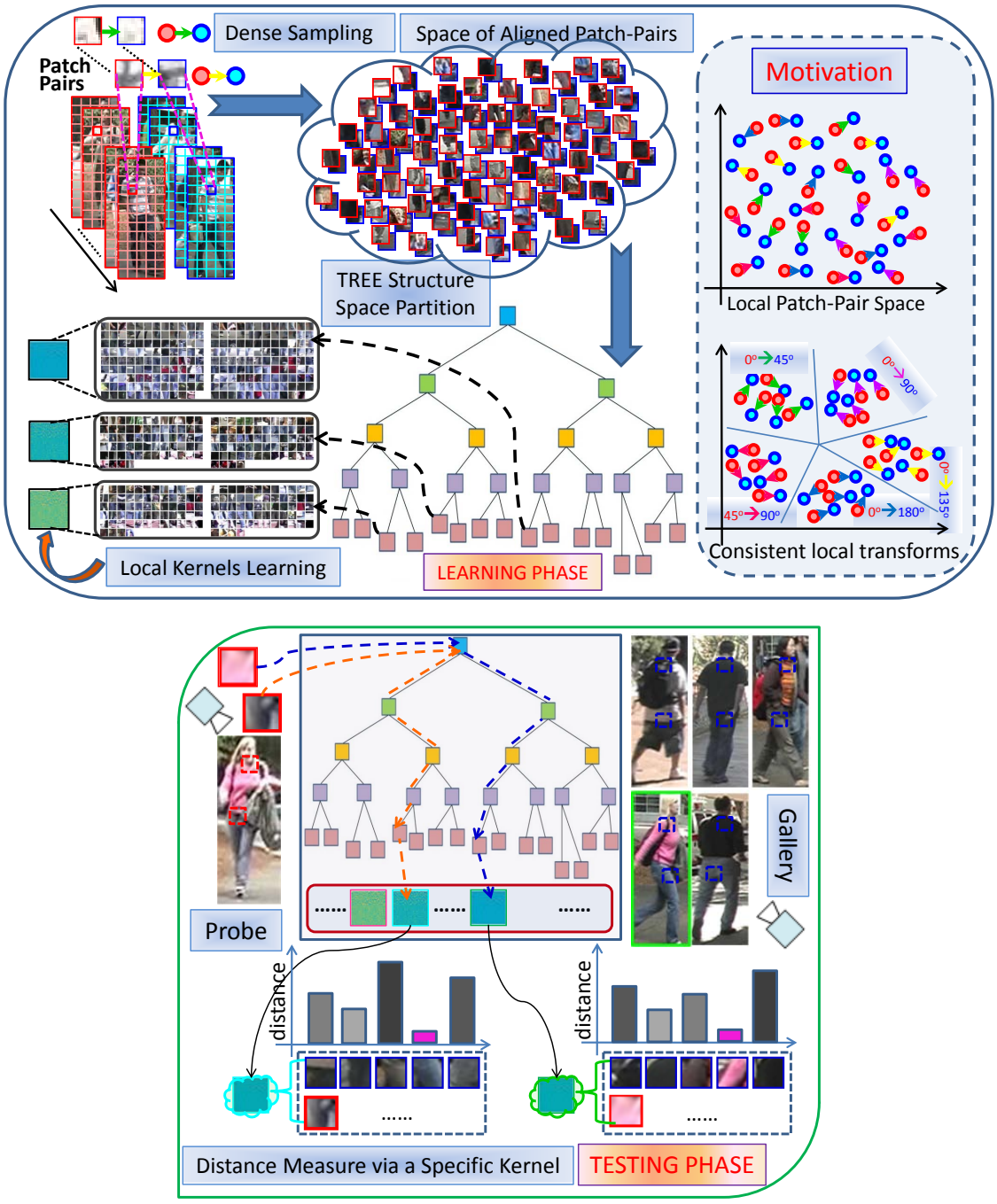

Figure 4.2: Illustration of the main idea. Top: learning phase, the aligned patch pairs of the same person from different cameras are separated in a tree structure based on the *consistent patch-to-patch transform criteria*. At each tree leaf, a simple but effective kernel is learned to describe the simplified transform. Bottom: testing phase, given a probe image, a suitable kernel will be selected based on the decision tree for each of its local image patch. With the optimal local kernel, the distance between the true patch pairs will be well minimized.

$\mathbf{y})^\top \mathbf{P}^\top \cdot \mathbf{P}(\mathbf{x}-\mathbf{y}) = (\mathbf{x}-\mathbf{y})^\top \mathbf{W}(\mathbf{x}-\mathbf{y})$, where $\mathbf{W} = \mathbf{P}^\top \cdot \mathbf{P}$ is a semi-definite matrix. However, as explained in Sec. 4.1, the complex transformation between cam$\mathcal{X}$ and cam$\mathcal{Y}$ is multi-modal, hence it cannot be well learned with a single fixed metric $\mathbf{W}$. In our work, we learn a more comprehensive overall mapping function $\mathcal{F}_{\mathbf{M}} : \mathcal{X} \rightarrow \mathcal{Y}$ which is parameterized by $\mathbf{M} = \{\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_K\}$, where each $\mathbf{m}_k$ represents a simple local transform that learned from a small set of ***automatically*** selected local patch-pairs in group $G_k = \{(x_i, y_i)_{i=1,2,...}^n\}$, from the training set of image pairs $\{\mathbf{X}, \mathbf{Y}\}$, where i is the subscript of each local patch in group $G_{k,k=1,2,...,K}$ and n denotes which image it comes from. Hereinafter, each independent local transform f_k parameterized by kernel $\mathbf{m_k}$ is denoted as $f_{\mathbf{m}_k}$. Learning such a kernel $\mathbf{m}_k$ is generally formulated using the empirical risk minimization:

$$\mathbf{m}_k^* = \arg\min_{\mathbf{m}_k} \frac{1}{|G_k|} \sum_{i \in G_k} \mathcal{L}(y_i, f_{\mathbf{m}_k}(x_i)) \tag{4.1}$$

Please note that each small group G_k is discovered automatically, we will illustrate how to partition the space of aligned local patch-pairs into different sub-spaces and get the resultant groups G_k in the next subsection via the decision tree structure. In this work, each $f_{\mathbf{m}_k}$ is defined as a linear mapping function describing the decomposed *uni-modal* local transform. The loss function $\mathcal{L}$ is simply defined as $(y_i - \mathbf{m}_k x_i)$, and $\mathbf{m}_k$ is just a linear mapping kernel that can be efficiently solved in closed form as $\hat{\mathbf{y}}\hat{\mathbf{x}}^\top(\hat{\mathbf{x}}\hat{\mathbf{x}}^\top + \lambda\mathbf{I})^{-1}$, where λ is a regularizing parameter being a small value, and $\hat{\mathbf{x}} = [x_1, x_2, ..., x_i, ..]_{x_i \in G_k}$ and $\hat{\mathbf{y}} = [y_1, y_2, ..., y_i, ..]_{y_i \in G_k}$. Finally, the overall inter-camera transformation from $\mathcal{X}$ to $\mathcal{Y}$ can be formulated as $\mathcal{F}_{\mathbf{M}} = \sum_1^K f_{\mathbf{m}_k}$, with each of the $f_{\mathbf{m}_k}$ representing one uni-modal transform that works on certain specific kind of image local patches.

Finally, our *local distance metric* is defined as $d_k(x_i, y_j) = ||y_j - f_{\mathbf{m}_k}(x_i)||^2$, where the optimal kernel $\mathbf{m}_k$ for each image patch x_i is *automatically* and discirminatively

assigned by the tree structure. Then, the *overall distance metric* is defined as:

$$\begin{aligned}\mathcal{D}(\mathbf{x},\mathbf{y}) &= ||\mathbf{y}-\mathcal{F}(\mathbf{x})||^2 \\ &= \textstyle\sum_{r,c}||y_{[r,c]} - \frac{1}{Q}\sum_q f^q_{\mathbf{m}_k}(x_{[r',c']})||^2\end{aligned} \tag{4.2}$$

where the subscript $[r,c]$ denotes the coordinates of each local patch in images $\mathbf{x},\mathbf{y}$. We use a greedy distance measure, which will be detailed in Sec. 4.2.3, to compute the pairwise patches distance, thus the $[r,c]$ in $\mathbf{y}$ and $[r',c']$ in $\mathbf{x}$ do not have to be identical. As to be introduced in Sec. 4.2.2, we formulate the uni-modal transform $f_{\mathbf{m}_k}(x)$ Q times in Q decision trees to avoid over-fitting, the final output thus is a mean value of the Q predictions.

4.2.2 Random Kernel Forest

A forest is an ensemble of Q decision trees T_q Criminisi and Shotton (2013). Given a sample x, the prediction of $T_q(x)$ from each tree is combined using an ensemble model, e.g., an average value, into a single output. Each decision tree consists of non-terminal (split) and terminal (leaf) nodes. A tree T_q classifies a sample $x \in \mathcal{X}$ by recursively branching left or right child node down the tree structure until reaching a leaf node. Each non-terminal node z in the tree is associated with a binary split function h with parameters θ_z:

$$h(\phi(x),\theta_z) = \begin{cases} 0 & \text{for} \quad \phi(x) < \tau_z \\ 1 & \text{for} \quad \phi(x) \geq \tau_z \end{cases} \tag{4.3}$$

Then, sample x will be sent to left if $h(\phi(x),\theta_z) = 0$, otherwise, right. The split function $h(\phi(x),\theta_z)$ can be arbitrarily complex, but a typical choice is just a threshold that a single entry on the feature vector x is compared to, e.g., $\theta_z = (k_z,\tau_z)$, then $h(\phi(x),\theta_z) = [x(k_z) < \tau_z]$, where $[\cdot]$ is an indicator function, and $x(k_z)$ is the k_z-th entry on the feature vector x. The function $\phi(x)$ also can be of other forms, for example, we use the "pairwise" difference of two entries on the feature vector x, i.e.,

$\phi(x) = x(k_1) - x(k_2)$. Both the two entries k_1, k_2 are randomly selected from feature vector x.

Suppose the training set, i.e., the aligned local patch-pairs $\mathcal{S} = \{(x_i, y_i)_{i=1,...}^{n=1,...,N}\}$, are extracted from the training image pairs $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ and $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N\}$. Training of the decision tree for joint spaces partition involves searching for the parameter θ_z of each split function $h(\phi(x), \theta_z)$, which can well split the training data to maximize an objective function, i.e., *information gain*.

$$\mathcal{I}_z = \mathcal{I}(\mathcal{S}_z, \mathcal{S}_z^L, \mathcal{S}_z^R) = E(\mathcal{S}_z) - \sum_{v \in L,R} \frac{|\mathcal{S}_z^v|}{|\mathcal{S}_z|} E(\mathcal{S}_z^v) \tag{4.4}$$

where $\mathcal{S}_z^L = \{(x, y) \in \mathcal{S}_z | h(\phi(x), \theta_z) = 0\}$, $\mathcal{S}_z^R = \mathcal{S}_z \setminus \mathcal{S}_z^L$, and the term E is an index function. Then, learning of the parameter θ_z is guided as to maximize $\mathcal{I}_z$. The same learning will be executed on each non-terminal nodes recursively until it reaches a leaf node or the gain falls below a threshold.

For typical classification problems, the term E is defined as the Shannon entropy $E(\mathcal{S}) = -\sum_c s_c \log(s_c)$, where s_c is the fraction of elements in $\mathcal{S}$ with label c Criminisi and Shotton (2013); Dollar and Zitnick (2013). In contrast, the index function E of our task is defined as the regression error $||\hat{\mathbf{y}}_z - \mathbf{m}_z \hat{\mathbf{x}}_z||^2$. Therefore, training for classification tasks partitions training samples into successive homogeneous sub-clusters, while tree training for our task partitions the local patch-pairs from two spaces jointly into successive sub-spaces where their inter-camera transforms become consistent and easier to formulate, layer by layer in the tree structure. Finally, we are able to define our local metric at each **leaf node** with a specific kernel, and the combination of those local kernels can approximate any complicated multi-modal inter-camera transformations. Some local kernels and their examplar local patches in the pair of subspaces are shown in Figure 4.3.

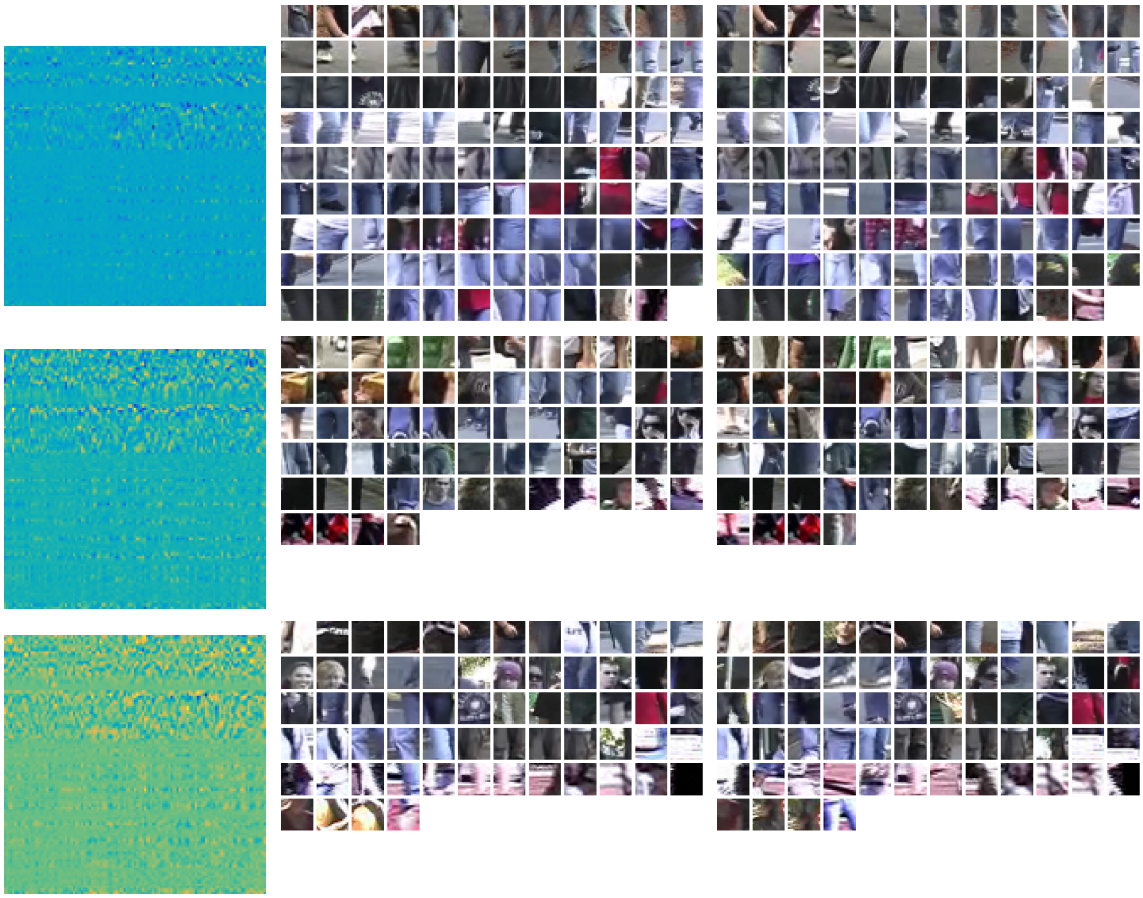

Figure 4.3: Some examplar local kernels and their learning patches from a pair of two subspaces (camS & camT) in the tree structure: *top*, node 224 at depth 8, *middle*, node 297 at depth 9, *bottom*, node 473 at depth 11. It is obvious that different local regions indicate different local metric kernels.

4.2.3 Patch Features and Alignment

Features of local patches: features of local patches on an overlapping dense grid are extracted, as shown in Figure 4.2. The features used for patch representation include: 10-bin color histogram extracted from each of the 3 channels of HSV color space and each of the 3 channels of LAB color space, 9-bin gradient histogram extracted from the intensity space, and 59-bin LBP features also extracted from the intensity space. The 8 channels of features are finally concatenated to form a final 128-dimensional feature vector for each local patch.

Constrained patches alignment: in each image $\mathbf{x}_n$, the appearance of human body is segmented into several horizontal stripes Pedagadi et al. (2013); Zhao et al. (2013b) to incorporate certain spatial constraint in patches matching and alignment. The feature of a local patch is denoted as $\{x^n_{r,c}\}$, indicating it's from the r-th row and c-th column on the dense grid. Since the two images from cam$\mathcal{X}$ and cam$\mathcal{Y}$ might be taken with different viewpoints, as shown in Figure 4.1, we need to roughly align the local patches in measure of the distance between two images $\mathbf{x}$ and $\mathbf{y}$. Therefore, when a local patch $\{x^n_{r,c}\}$ is matched to a corresponding one in the image $\mathbf{y}_u : \{y^u_{r,c}\}$, its search is constrained to the set $\{y^u_{[r-1,r+1],c=1,\ldots,C}\}$. With the searching in a small range $[r-1, r+1]$, we can relieve the neg-effect in patches matching caused by the vertical misalignment. We perform the patch matching in a greedy way, in both the extraction of training patch-pairs and the testing of images re-identification. Each patch $x^n_{r,c}$ is matched to its nearest neighbor in its searching set $\{y^u_{[r-1,r+1],c=1,\ldots,C}\}$, then the corresponding pair in the set will be removed in next iteration, as illustrated in Figure 4.4. Finally, each local patch in image $\mathbf{x}_n$ is aligned to a unique one in image $\mathbf{y}_u$. Then, the distance between the two images is the summation of all the pairwise distance of each two local patches from $\mathbf{x}_n$, $\mathbf{y}_u$, as in Eq. 4.2. The retrieved image in gallery for the query image is the one gives the smallest distance value $\mathcal{D}(\mathbf{x}_n, \mathbf{y}_u)$.

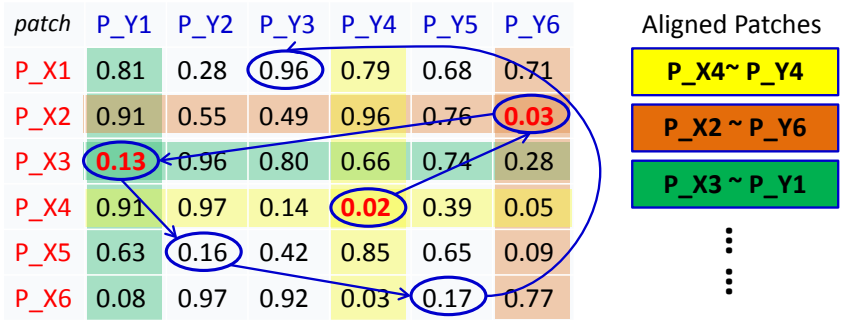

Figure 4.4: Illustration of the greedy local patches matching via pairwise distance. Suppose 6×2 patches are doing matching from two vertical strips of $\mathbf{x}, \mathbf{y}$, the sequence of the matched patches in this example are denoted as in color yellow, orange and green ... .

4.2.4 Discussion and Implementation

If viewed from the perspective of motivation, our work is most close to the LAFT approach Li and Wang (2013), which jointly partitions the image spaces of two camera views into different subspaces according to the similarity of inter-camera transforms. However, the main difference between our works are that: (i) LAFT partitions the image space of each camera view instead of the more fine local patches space, where the problem of local regions suffering from different transforms can be better tackled; (ii) LAFT uses a gating network to softly assign the given image pair to a configuration type and requires feature selection with sparsity and log-determinant divergence regularization. In contrast, we assign the optimal kernel to the given local patch much more efficiently in the tree structure and do not require post feature selection. Viewed from the perspective of methodology, our work is also close to Liu et al. (2014); Zhao et al. (2014) as we all play with local patches. However, the assumption in the dictionary learning based approach Liu et al. (2014) that the local patches in different camera views have similar manifold structure is obvious not solid enough as the configuration are multi-modal. Two similar patches in one camera might correspond to two totally different patches in another camera. Zhao Zhao et al. (2014) partitioned the local patches space into sub-clusters based on patch features

from one single camera view. Therefore, the cross-view transform in each sub-cluster is still multi-modal, the mid-filters learned in these sub-clusters might not be able to respond to the cross-view invariant features thoroughly.

Training details: training of the decision tree plays the main role in learning of the overall inter-camera transformation $\mathcal{F}$. Random forest prevents over-fitting by training multiple de-correlated trees and combining their outputs. To achieve sufficient diversity of trees, we trained 20 trees in the forest. To learn the parameter θ_z at each non-terminal node z in training of each tree, we randomly sub-sample 1024 patch-pairs, 20 pairs of the entry on the feature vector and take 10 random guesses for the threshold τ_z. The decision tree terminates split and creates a leaf once the number of patch pairs is less than 128.

Processing Flow: the pseudo-code of the overall flow for learning the random kernel forest and testing for a query are also shown as in Algorithm 3.

4.3 Experiments

4.3.1 Datasets and Protocols

We conduct experiments on three most frequently used datasets: the "viewpoint invariant pedestrian recognition dataset" (VIPeR) Gray and Tao (2008), the "QMUL underround re-identification dataset" (GRID) Loy and Tang (2009) and the "CUHK person re-identification dataset" (CUHK01) Cheng et al. (2011). All three datasets are very challenging for re-id problems due to the significant variations in viewpoints, poses, illuminations, and also their low image resolutions with occlusions and different backgrounds.

VIPeR: it contains 632 pedestrian image pairs that captured by two hand-carried cameras in outdoor environment. All the images are scaled to the same size of 128×48 for evaluation. Each pair contains two images of the same person observed from

Algorithm 3 Processing flow of kernel forest learning and testing

Training Phase:

Input: $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$: training images in cam$\mathcal{X}$.
$\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N\}$: training images in cam$\mathcal{Y}$.
υ: number of samples allowed for split.
ρ: maximum levels of the tree.

Output: A random kernel forest of Q decision trees T_q.

Extract a set of local patch pairs $\mathcal{S}_{z=0}\{(x_i, y_i)_{i=1,2,...}\}$ based on the constrained greedy patches matching from $\mathbf{X}, \mathbf{Y}$.
while $q \leq Q$ **do**
  Train each decision tree T_q independently.
  while $|\mathcal{S}_z| \geq \upsilon$ **do**
    node parameter learning: select parameter θ_z that maximize $\mathcal{I}_z$ in Eq. 4.4.
    data split: send x to left if $h(\phi(x), \theta_z) = 0$,
      otherwise, right, as in Eq. 4.3.
  end while
  Create leaf node with a local linear kernel $\mathbf{m}_k$.
end while

Testing Phase:

Input: query image $\mathbf{x}$ from cam$\mathcal{X}$.
a set of gallery images $\mathbf{Y}_g = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N\}$.

Output: $\mathbf{y}^*_\mathbf{n}$: the same person as in $\mathbf{x}$.

dense sample local patches $(x_i)_{i=1,...,\mathbf{r}*\mathbf{c}}$ in image $\mathbf{x}$.
while $q \leq Q$ **do**
  while $i \leq \mathbf{r} * \mathbf{c}$ **do**
    Input patch x_i into tree T_q, and reach leaf node z.
    Calculate the mapping $f_{\mathbf{m}_k}(x_i)$ with the local kernel $\mathbf{m}_k$ associated on the leaf node z.
  end while
end while
Calculate the averaged mapping of each local patch x_i as $\frac{1}{Q}\sum_q f^q_{\mathbf{m}_k}(x_i)$.
Calculate the distance between two images $\mathbf{x}$ and $\mathbf{y}_n$ as in Eq. 4.2, and choose the image $\mathbf{y}^*_\mathbf{n}$.

two camera views with pose changes (mostly $> 90^o$ degree) and different lighting conditions.

GRID: it contains 250 pedestrian image pairs that captured from 8 disjoint camera views installed in a busy underground station. All the images are scaled to the same size of 300×100 for evaluation. Each pair contains two images of the same individual seen from different camera views. Except for the common challenges (pose changes, etc), the gallery set also contains 775 distracting images which do not match any person in the probe set, bringing much more difficulty in re-identification for a probe (query) image.

CUHK01: this is a multi-shot dataset containing 971 pedestrians captured from two disjoint camera views, with 2 images per person in each view. All the images are scaled to the same size of 160×60. As it contains much more instances, it has been used for evaluation of deep learning approaches.

Protocals: The pedestrians in each dataset are separated into the training set and the testing set, such that each person appears only once in either the training set or the testing set. The testing set is also partitioned into two sets: the probe set and the gallery set. For the VIPeR dataset, the images in camera A are used as probe images, and the images in camera B are used as gallery images. The GRID dataset already defined the probe set and the gallery set, with 775 distracting images added in the gallery set. For the CUHK01 dataset, the first 2 images of each person are used as probe images and the latter 2 images from another view are stored in the gallery set. According to the existing works in literature, the performances are reported quantitatively as the standard Cumulated Matching Characteristics (CMC) curves, and the performance is the averaged results of ten trials. In CMC curves, the Rank-κ matching rate is the rate of correct match at rank κ, and the cumulated values of recognition rate at all ranks is recorded as the CMC curve. The parameters in learning of the random kernel forest are illustrated in Sec. 4.2.4. For dense local patches sampling of the images in each dataset, 15×5, 24×8, 19×6 overlapping local patches are extracted in VIPeR, GRID, CUHK01, respectively.

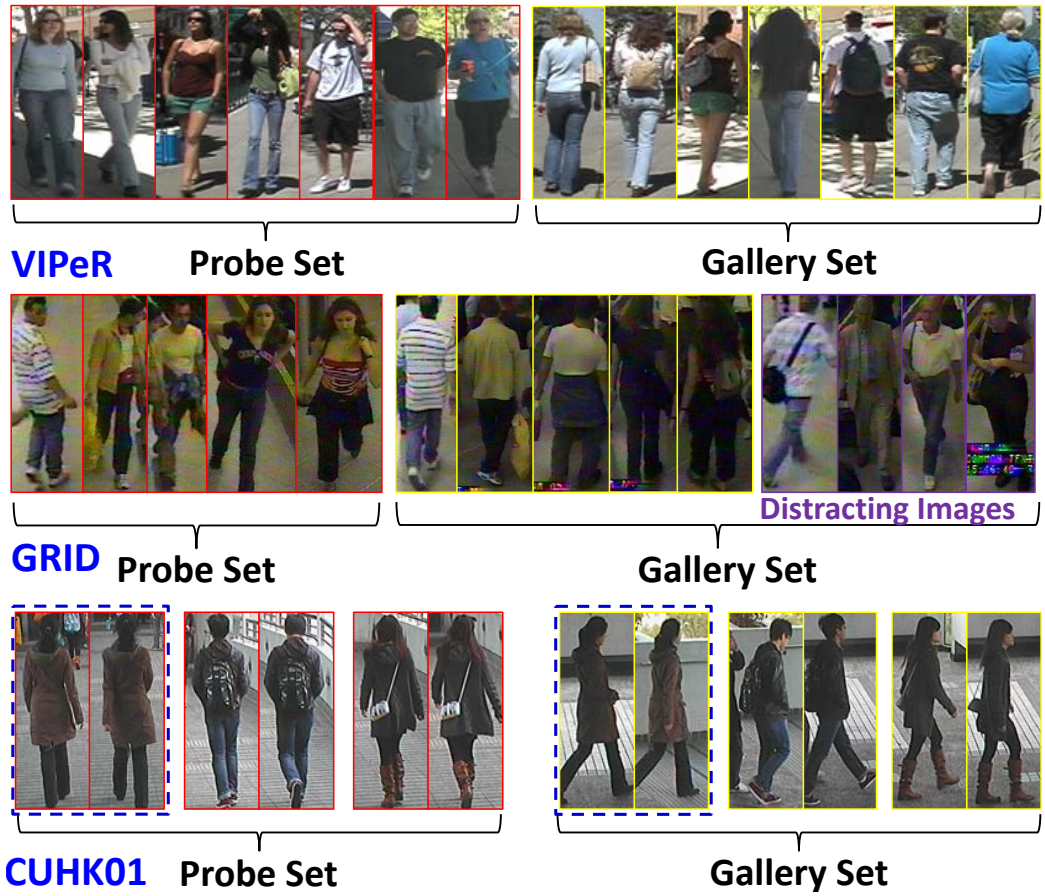

Figure 4.5: Examplar image pairs in probe set and gallery set from datasets of VIPeR (top), GRID, CUHK01(bottom), respectively.

Figure 4.6: Left-4: similarity distribution of local regions in matching. Right-2: spatial distribution of 127 local kernels in an example image.

4.3.2 Empirical Analysis

We investigate how some of the terms in our random kernel forest influence the final re-identification performance. All the analysis and evaluations in this sub-section are based on the VIPeR dataset.

Effect of local kernels: The distance between the query image and each gallery image is the summation of all the pairwise distances of local patches. To tell which local regions contribute the most to discriminate the correct match in the gallery set, we show the similarity distribution of one example image pair in left of Figure 4.6, from which we can find that these discriminative regions mostly focus on human body parts. In addition, we also show the spatial distributions of the local kernels on one example image in right of Figure 4.6, which also demonstrates our hypothesis that the inter-camera transforms at different local regions indeed vary accordingly.

Forest diversity: The diversity of trees in the kernel forest is crucial in traditional random forest classifiers. In fact, the accuracy of each single tree is sacrificed in favor of a highly diverse ensemble. Therefore, we vary the number of trees Q in the forest and check their influence on the final re-id performance. As shown by the results in Figure 4.7 (a), a larger number of trees produced higher re-id performance. However, once the number of trees is large enough, the performance becomes stable. Based on the empirical study, we choose the number of trees in our forest as 20, which is relatively small while producing good performance.

Partition of image space and patch space: As mentioned in Sec. 4.2.4, based on similar motivation that finding a subspace where the cross-view data pair inside have consistent transform, the LAFT Li and Wang (2013) partitions the image space while ours partition the more fine local patch space. We thus conduct two tests for evaluation based on the VIPeR dataset, one uses 316 persons in training set (316 gallery images in test) and the other uses only 100 persons in training, resulting in 532 gallery images in test. The performance comparison between the two approaches are shown in Figure 4.7 (b). We can observe that in the first test, our approach performs

better in the range of a small κ (rank 2-15), while in the second more challenging test with much less training samples and a larger gallery set, our performance is obviously much better than the LAFT approach.

4.3.3 Quantitative Evaluation

In this subsection, we compare our approach to the other existing works on several standard datasets for evaluation.

VIPeR: two protocals were defined for evaluation on this dataset: the first one randomly selects 316 persons to form the training set and results in 316 persons in testing set; the other one randomly selects 100 persons to form the training set and results in 532 persons in test. Our approach is compared to the other existing works including: SDALF Farenzena et al. (2010), LF Pedagadi et al. (2013), SSCDL Liu et al. (2014), SalMat Zhao et al. (2013b), IRS Lisanti et al. (2015), RPLM Hirzer et al. (2012), mFilter Zhao et al. (2014), QALF Zheng et al. (2015), LADF Li et al. (2013b), PCCA Mignon and Jurie (2012), MtMCML Ma et al. (2014), MFA Xiong et al. (2014), LAFT Li and Wang (2013), kLFDA Xiong et al. (2014), ReML Chen et al. (2015b). The performance comparison is shown in Figure 4.7 (c) and (d) by CMC curves. From these results, we can find that our approach gives the second best performance in the first test and the best performance in the second test. We also summarize the performance comparison in Tables 4.1&4.2 to show the matching rate values more straightforwardly. It is clear that our approach achieves 29.1% and 16.0% rank-1 matching rate in the two tests, which is very competitive compared to the other results in literature. The rank-20 matching rate for our approach is 83.8% and 67.4% in the two tests, which also outperform most of the other methods.

GRID: experiments on this dataset were conducted according to the 10 data partitions provided along with the dataset. In each partition, the image pairs from 125 randomly selected individuals are used for training, and the rest 125 persons together with the 775 irrelevant distracting images form the gallery set in test. Our

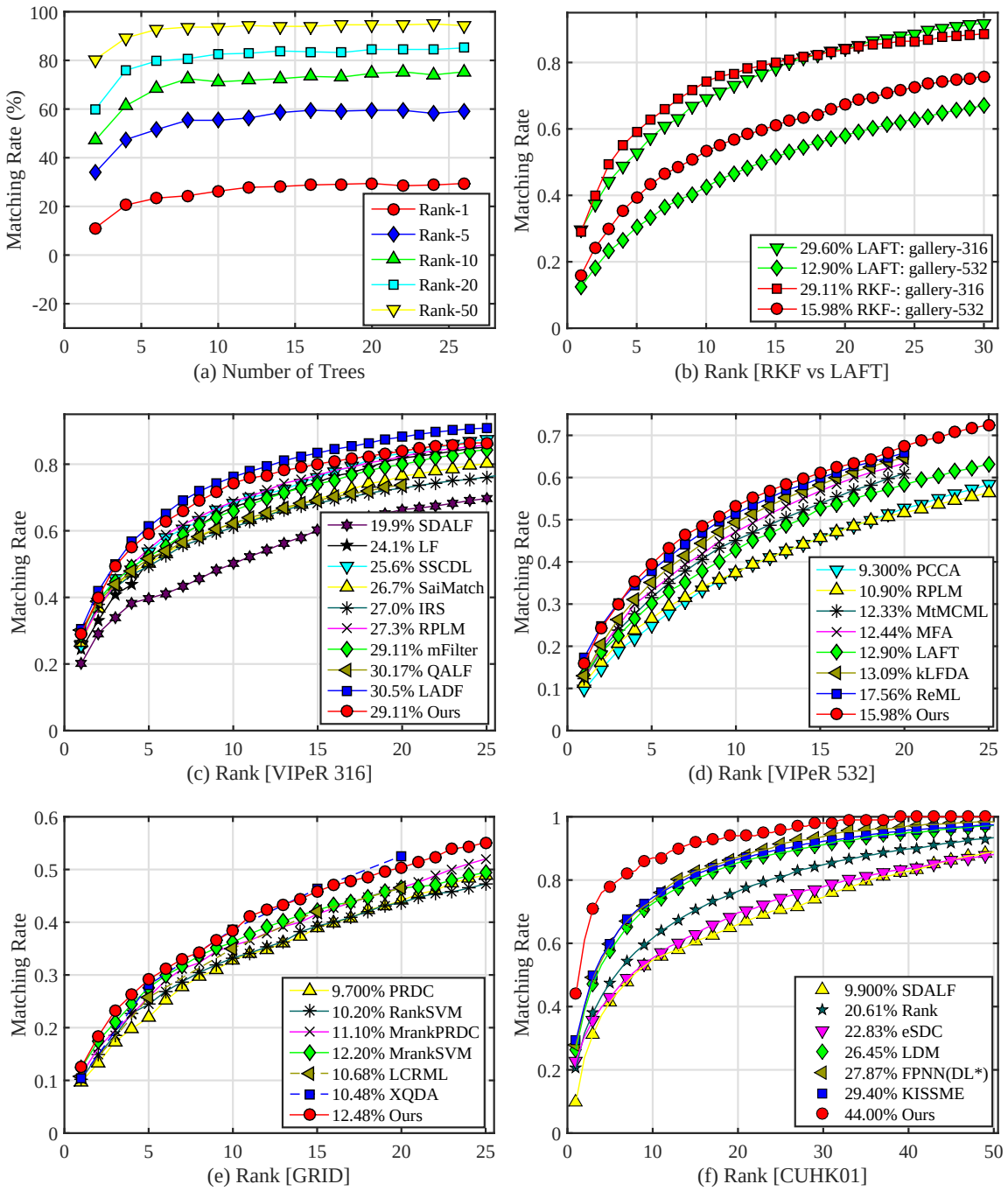

Figure 4.7: Evaluations: (a) Performance comparison of different numbers of trees in random kernel forest. (b) Comparison between RKF and LAFT via CMC curves. (c) CMC curves on VIPeR dataset with 316 gallery images. (d) CMC curves on VIPeR dataset with 532 gallery images. (e) CMC curves on GRID dataset with 900 gallery images. (f) CMC curves on CUHK01 dataset with 100 gallery images.

Table 4.1: Top ranked matching rates (%) on VIPeR dataset with 316 gallery images.

Method	$\kappa = 1$	$\kappa = 5$	$\kappa = 10$	$\kappa = 20$	ref
PRDC	15.7	38.4	53.9	70.1	CVPR11 Zheng et al. (2011)
PCCA	19.3	51.2	64.9	77.5	CVPR12 Mignon and Jurie (2012)
KISSME	19.6	48.2	62.2	76.9	CVPR12 Kostinger et al. (2012)
SDALF	19.9	38.9	49.4	65.7	CVPR10 Farenzena et al. (2010)
eLDFA	22.3	47.0	60.0	71.0	ECCV12 Ma et al. (2012b)
LF	24.1	51.2	67.1	82.0	CVPR13 Pedagadi et al. (2013)
SSCDL	25.6	53.7	68.1	83.6	CVPR14 Liu et al. (2014)
SalMat	26.7	50.7	62.4	76.4	CVPR13 Zhao et al. (2013b)
IRS	27.0	49.4	61.1	72.8	PAMI15 Lisanti et al. (2015)
RPLM	27.3	54.5	68.8	82.4	ECCV12 Hirzer et al. (2012)
mFilter	29.1	52.3	66.0	79.9	CVPR14 Zhao et al. (2014)
QALF	30.2	51.6	62.4	73.8	CVPR15 Zheng et al. (2015)
LADF	30.5	61.2	76.2	88.2	CVPR13 Li et al. (2013b)
Ours	29.1	59.2	74.4	83.8	1st, 2nd, 3rd

approach is compared to some recently published results: PRDC Zheng et al. (2011), RankSVM Prosser et al. (2010), MrankPRDC Loy et al. (2013), MrankSVM Loy et al. (2013), LCRML J. Chen and Wang (2014), XQDA Liao et al. (2015) in Figure 4.7 (e) and Table 4.3. The CMC curves and top rank matching rates show our approach achieves very competitive results on this benchmark.

CUHK01: this is a multi-shot dataset containing 971 persons. 100 persons are randomly selected in test, and the rest 871 persons are used for training. This protocol was designed for deep learning in FPNN Li et al. (2014b). Figure 4.7 (f) and Table 4.4 compared the performance of our approach to the other existing works including FPNN, eSDC Zhao et al. (2013b), KISSME Kostinger et al. (2012), LDM Guillaumin et al. (2009), etc. The results show that our approach outperforms the other existing works by a wide margin ($>$ 15% than FPNN), with the rank-1 matching rate being 44%. This might attribute to the fact that deep leaning approaches usually requires a large amount of training data, otherwise the network tend to over-fitting. In summary, all the above results also show that our approach is able to achieve very competitive performance without the strict requirements on training data as in deep learning.

Table 4.2: Top ranked matching rates (%) on VIPeR dataset with 532 gallery images.

Method	$\kappa = 1$	$\kappa = 5$	$\kappa = 10$	$\kappa = 20$	ref
PCCA	9.3	24.9	37.4	52.9	CVPR12 Mignon and Jurie (2012)
RPLM	10.9	26.7	37.7	51.6	ECCV12 Hirzer et al. (2012)
MtMCML	12.3	31.6	45.1	61.1	TIP14 Ma et al. (2014)
MFA	12.4	33.3	47.2	63.5	ECCV14 Xiong et al. (2014)
LAFT	12.9	30.3	42.7	58.0	CVPR13 Li and Wang (2013)
kLFDA	13.1	35.2	49.4	65.0	ECCV14 Xiong et al. (2014)
ReML	17.5	37.9	51.8	66.0	TIP15 Chen et al. (2015b)
Ours	16.0	39.5	53.3	67.4	1st, 2nd, 3rd

Table 4.3: Top ranked matching rates (%) on GRID dataset with 900 gallery images.

Method	$\kappa = 1$	$\kappa = 5$	$\kappa = 10$	$\kappa = 20$	ref
PRDC	9.7	22.0	33.0	44.3	CVPR11 Zheng et al. (2011)
RankSVM	10.2	24.6	33.3	43.7	BMVC10 Prosser et al. (2010)
MrankPRDC	11.1	26.1	35.8	46.6	ICIP13 Loy et al. (2013)
MrankSVM	12.2	27.8	36.3	46.6	ICIP13 Loy et al. (2013)
LCRML	10.7	25.8	35.0	46.5	ICPR14 J. Chen and Wang (2014)
XQDA	10.5	28.1	38.6	52.6	CVPR15 Liao et al. (2015)
Ours	12.5	29.2	38.3	50.3	1st, 2nd, 3rd

Table 4.4: Top ranked matching rates (%) on CUHK01 with 100 gallery images.

Method	$\kappa = 1$	$\kappa = 5$	$\kappa = 10$	$\kappa = 20$	ref
SDALF	9.9	41.5	54.7	66.0	CVPR10 Farenzena et al. (2010)
Rank	20.6	47.6	61.6	76.5	ICML10 Mcfee and Lanckriet (2010)
eSDC	22.8	43.0	55.3	69.7	CVPR13 Zhao et al. (2013b)
LDM	26.5	57.6	72.6	85.5	ICCV09 Guillaumin et al. (2009)
FPNN	27.9	59.7	73.4	87.3	CVPR14 Li et al. (2014b)
KISSME	29.4	59.8	74.5	86.6	CVPR12 Kostinger et al. (2012)
Ours	44.0	78.5	86.7	94.0	1st, 2nd, 3rd

4.4 Summary

This work presented a novel approach based on the random kernel forest for person re-identification across disjoint camera views with complicated appearance variations. The complex inter-camera transformation is modelled by a combination of many local functions, which formulate each local transform in a much simpler but effective manner. Both the decomposition of the overall inter-camera transformation and the local metric kernels for re-identification are discovered automatically by the aligned local training patch-pairs using the random forest framework. Any local patch in a query image is assigned a specific kernel in the tree structure, then the local metric is able to generate a minimized distance between the true patch-pairs. Extensive experimental results showed that the proposed random kernel forest achieved very competitive re-identification performance as compared to the existing works.

Chapter 5

Multi-Event Detection and Recognition in Smart Grid

Event analysis has been an important component in any situational awareness system. However, most state-of-the-art techniques can only handle single event analysis. This chapter tackles the challenging problem of multi-event detection and recognition in smart grid system. We propose a novel conceptual framework, referred to as *event unmixing*, where we consider real-world disturbances are mixtures of more than one constituent root events, which are also transferable across the data domains of single event and multi-event. This concept is a key enabler for analysis of multi-event to go beyond what are immediately detectable in a signal, providing high-resolution data understanding at a much finer scale. We interpret the event formation process from a linear mixing perspective and propose an innovative nonnegative and sparse event unmixing (NSEU) approach for multiple event separation and temporal identification. Experimental results demonstrate that the framework is quite reliable to detect and recognize multiple constituent cascading events as well as identify their occurring time with high accuracy.

5.1 Introduction

The US electric power system provides vital links that achieve essential continuity of service from generating plants to end users. However, the ever-increasing complexity in sensing and actuation, compounded by limited knowledge of the accurate system status have resulted in major system failures, such as the massive power blackout of Aug. 2003 and the most recent Arizona/California blackout of Sep.2011. Therefore, it becomes essential that the wide-area situational awareness (WASA) systems enables monitoring of bulk power systems and provide critical information for understanding and responding to power system disturbances and cascading blackouts.

The frequency disturbance recorder (FDR) sensor is able to collect instantaneous information of voltage phasor and frequency at a low-voltage distribution level using ordinary 120-V wall outlets. An US-wide Frequency Monitoring Network (FNET) has thus been implemented (Zhong et al., 2005; Liu, 2006; Gardener and Liu, 2007) based on these low-cost sensors and serves the entire north American power grid through advanced situational awareness techniques. When an event occurs in a power grid, the imbalance between power generation and load consumption causes sudden frequency changes within the system that can be used as a good indicator for event analysis. A couple of works have been reported for conducting event analysis using data collected from the FNET (Zhang et al., 2010; Li et al., 2010; Zhao et al., 2008; Xia et al., 2007; Gardner et al., 2006; Kook and Liu, 2011). Although successful, these state-of-the-art techniques only handle disturbances caused by single event. If multiple cascading events are involved, existing techniques can only detect frequency disturbances caused by the initial event, and the frequency disturbances from successive events might be overshadowed by the continued frequency fluctuation from the initial event. We also observe that when multiple events occur in cascading fashion, the electromechanical waves generated will interfere with each other, and the measurement taken at a FDR sensor would more than likely be a "mixture" of multiple constituent event signals. Therefore, how to determine the number of constituent events that occurred and the

categories of the events with precise estimation of the occurring time using simply the observed frequency signal presents a very challenging problem.

In this chapter, we study the problem of multi-event detection and recognition with an assumption that we are given a database with a set of signal recordings from single event. However, because of the correlation between electrical devices in a smart grid system, the signal of multi-event will generate specific correlation characteristics, resulting in variations to simple combination of the single events. Moreover, learning samples of multi-event signal with ground truth for doing regression requires a large number of multi-event signals, which is hard to obtain. Thus, how to make use of the knowledge that extracted from the collection of single event signals poses challenge for this cross domain analysis. In this work, the main contributions are three-fold: 1) the formulation of the multi-event analysis problem using a linear mixing model based on commonly shared root-patterns across data domains, 2) the construction of the signature dictionary that incorporates temporal information subtly to reflect the event dynamics in power grid, and 3) the validation of the proposed approaches using extensive simulations and real case studies.

5.1.1 Problem Formulation

Similar to the ubiquitous existence of mixed measurements, events might not occur in a pure and isolated fashion, especially in power grid. Taking the major U.S. western blackout in 1996 as an example, at the very beginning of the blackout, two parallel lines were tripped due to a fault and mis-operation of the protective equipments, and consequently some generation was tripped as a correct special protection system (SPS) response. Then, the third line was disconnected due to bad connectors in a distance relay. After more than 20 seconds of these disturbances, the last straw of the collapse occurred when the Mill Creek - Antelope line tripped due to an undesired operation of a protective relay. After this line trip, the system collapsed within three

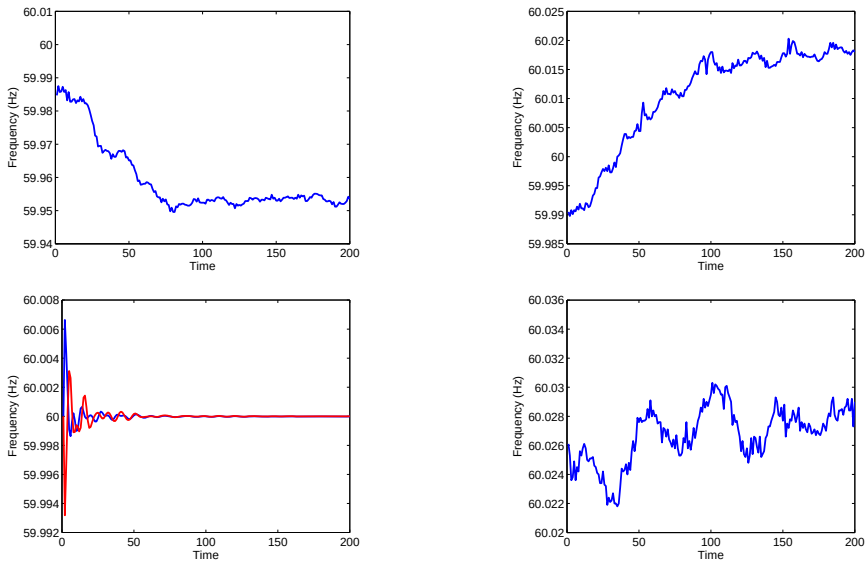

Figure 5.1: Four types typical root events: generator trip, load shedding, line trips, oscillation.

seconds. Therefore, to be able to provide high-resolution understanding of the power system dynamics, it is essential to perform multi-event analysis.

Typical disturbances in smart grid system fall into one of four categories, including generator trip (GT), load shedding (LS), line trip (LT), and oscillation (OS), which we refer to as the "root events" or "pure events". Figure 5.1 displays the frequency variation patterns when each of these root events occurs. Denote the frequency signal collected at a FDR as $\mathbf{x}$ (also referred to as the *measured signal*) and the frequency variation pattern of each root event as $\mathbf{s}$ (also referred to as the *source signal*). We propose a new conceptual framework for the study of multi-event analysis, where we consider the disturbances sensed at each FDR, $\mathbf{x}$, as a linear mixture of a limited number of root events $\mathbf{s}$. The formulation can thus be expressed as below:

$$\mathbf{x} = \mathbf{S}\alpha + \epsilon \tag{5.1}$$

where $\mathbf{x} \in \mathbb{R}^l$ is the observation and l is the number of measurements corresponding to the time over which an event is measured. $\mathbf{S} \in \mathbb{R}^{l \times c}$ is the root event signature

dictionary whose columns, $\{\mathbf{s}_j\}_{j=1}^{c} \in \mathbb{R}^l$, correspond to the different pattern profiles of the root events, and α is the mixing coefficient vector. The possible error and noises are taken into account by the l-dimensional column vector ϵ. This concept is the key to accurate event analysis that goes beyond immediately detectable information in an observation and uncovers the true cause(s) of the multi-cascading-event. To realize this framework, however, we have to answer the following questions:

1. Is it valid to use a *linear* mixing model to formulate the mixture observation sensed at a FDR?
2. How can one obtain the profiles of root events given the complexity and dynamic nature of the power grid system?
3. How can one incorporate the different starting times of cascading root events in the construction of $\mathbf{S}$?
4. How can one solve α for event detection and recognition purposes in an on-line fashion?

We defer the discussions of the first issue to Sec. 5.2.1 and the second issue to Sec. 5.2.2. To resolve the other two issues, we construct an over-determined signature dictionary to incorporate pattern profiles of various types of root events as well as temporal information on how the events cascade. In this way we can simultaneously detect events type and events starting time in one unmixing procedure. In addition, since the number of events is much less than the number of pattern profiles in the dictionary, the "sparsity" enforced on the coefficient vector α becomes an appropriate constraint that not only reduces the solution space, making the problem well-posed, but it also helps in identification of event type and precise starting time. The sparse representation of the measured mixture is achieved through solving an ℓ^1-regularized least squares problem, which can be done efficiently through as a Lasso problem.

To further improve the robustness, we also enforce the "nonnegativity" constraint onto the coefficient vector α. There are two reasons for adding this constraint: first, from the energy perspective, physical laws govern the power flow in a grid system

where the mixture of electromechanical waves generated by multiple disturbances should be only an *additive* combination of the constituent components; second, since generator trips generally cause a decrease in frequency while load sheddings generally cause an increase in frequency, the nonnegativity constraint on α is especially helpful to eliminate error cases where a load shedding event is mistakenly recognized as a generator trip in the reverse way.

Based on the aforementioned discussion, we mathematically formulate the event unmixing problem as below:

$$\min \|\alpha\|_0 \text{ s.t. } \|\mathbf{x} - \mathbf{S}\alpha\|_2^2 \leq \epsilon,\ \alpha \geq 0 \tag{5.2}$$

where $\|\alpha\|_0$ represents ℓ^0-norm, which is defined as the number of non-zero entries in vector α. The proposed cost function consists of two components with one measuring approximation fidelity of the linear mixing model (i.e., $\|\mathbf{x} - \mathbf{S}\alpha\|_2^2$) and the other measuring sparsity of the coefficient vector, α. We refer to the proposed approach as nonnegative sparse event unmixing (NSEU) for detection, recognition, and temporal localization of multiple cascading events in power grid. In the following sections, we will discuss solutions to the four questions raised in this section.

5.2 Methodology

5.2.1 Linear Mixing Model

A power grid disturbance, in many ways, exhibits characteristics of electromechanical wave propagation phenomenon. When electromechanical waves from different sources interfere with each other, the standard wave equation is nonlinear (Thorp et al., 1998), which is difficult for explicit calculation; however, it is much more simple to analyse the power variation in a grid system if it is expressed as a frequency signal, as the

generation/load loss is proportional to the frequency variation (Dong et al., 2007):

$$\Delta P = \beta \Delta f \tag{5.3}$$

where β is a frequency sensitive constant (Bykhovsky and Chow, 2003). Based on Eq. 5.3, it will be reliable to approximate the mixture of frequency signals as a linear addition of frequency signals caused by several root constituent events.

5.2.2 Signature Dictionary Construction

Both the second and the third questions raised in Sec. 5.1.1 regarding the formulation of event unmixing are related to construction of the root event signature dictionary $\mathbf{S}$. In our work, we hypothesis the root event patterns are transferable between the data domains of single event and multi-event, thus these commonly shared patterns can be extracted from the collection of single event. The dictionary construction involves two steps. First, the root event patterns are learned from the training data previously collected from FDRs recording disturbances experienced during single events. Second, the temporal information (i.e., event starting time) is embedded into dictionary $\mathbf{S}$ by augmentation and padding operations with the learned root event patterns.

Learning Root Event Patterns: As mentioned in Sec. 5.1.1, typical power system disturbances (events) fall into one of four categories, including generator trip (GT), load shedding (LS, or load drop), line trip (LT), oscillation (OS). As described in (Zhao et al., 2008; Qi et al., 2011; Markham and Liu, 2011), events of the same category generally share similar characteristics. For example, a generation trip always starts with a rapid frequency drop and a load shedding always starts with a frequency increase. Meanwhile, events of the same category also might contain certain degree of differences because of the different setups of intrinsic parameters, including power flow output, consumption, etc. It is impractical to include every single event pattern in the dictionary which would affect the performance of online processing, especially for large-scale systems, i.e., the power grid. Meanwhile, these patterns are attached to

each unique devices thus not flexible enough to be taken as latent root patterns. We, instead, resort to machine learning to find a set of representative root event patterns for each category. Hereinafter, we refer to these root event patterns as "*root-patterns*".

Various methods can be adopted for extracting the root-patterns, including, for example, K-means clustering and dictionary learning (Sprechmann and Sapiro, 2010b; Ophir et al., 2011). The K-means clustering is chosen to perform the root-pattern extraction task. Given a set of single event observations $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ of the same event category collected from either the recordings of a real system (e.g., FNET) or simulation, K-means clustering aims to partition the n observations into K subsets, so as to minimize the within-cluster sum-of-squares error:

$$arg \min_{\Phi_E} \sum_{i=1}^{K} \sum_{\mathbf{v}_j \in E_i} \|\mathbf{v}_j - \mathbf{e}_i\|^2 \tag{5.4}$$

where $\Phi_E = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_K\}$ refers to the set of extracted root-patterns and $\mathbf{e}_i$ is the centroid of each cluster E_i. Notice that before performing K-means clustering on $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$, each vector should be normalized to have unit ℓ^2-norm, such that the scale ambiguity (Yang et al., 2010a; Ji et al., 2009) in learning root-patterns can be eliminated. In addition, when performing K-means clustering, instead of directly using the mean vectors as cluster centroids, we use the nearest observation $\mathbf{v}_i$ as the centroid for the ith cluster in the final loop. In this way, the learned root-patterns are also closer to real data.

In this chapter, we expect to detect and recognize three different types of events, i.e., GT, LS, and LT. We set K=6 for K-means clustering, then the number of root-patterns is $3 \times 6 = 18$.

Incorporating the Temporal Information: After learning the set of root-patterns from each category, the next step is to incorporate the temporal information into construction of the root event signature dictionary, $\mathbf{S}$. We refer to each column of $\mathbf{S}$ that already embeds temporal information as the "*temporal root-pattern*". The reason for doing this is that any event, being single or consisting of cascading events,

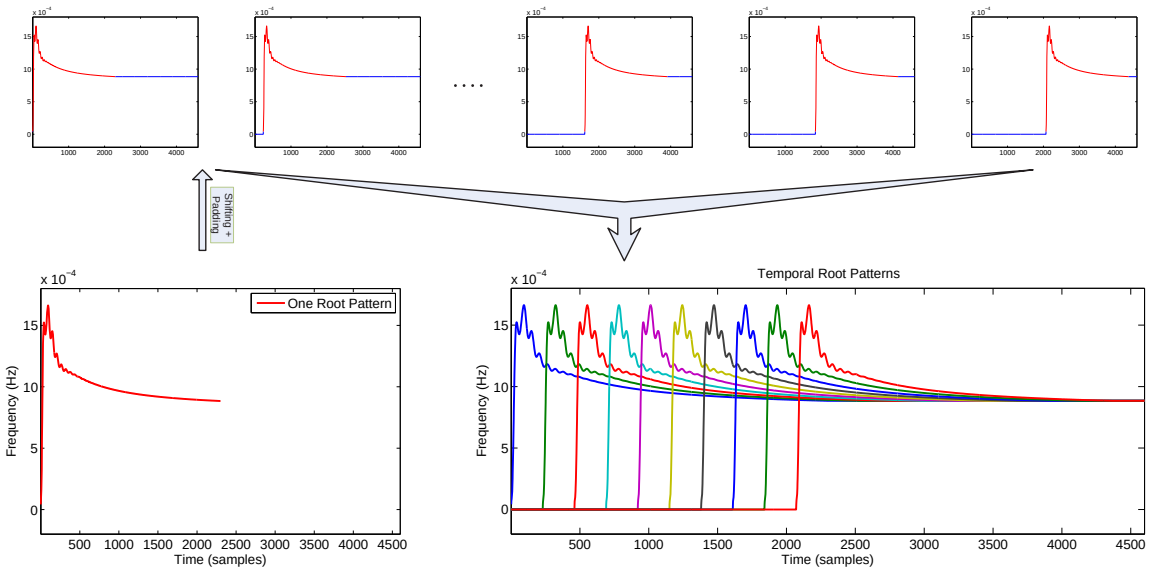

Figure 5.2: Shifting and padding one root-pattern to be "temporal root-patterns", bottom-left: one root-pattern, top: temporal root-patterns of different starting time, bottom-right: a group of temporal root-patterns to be incorporated in dictionary S.

can start at any time and also can last for a different period of time. To make sure that the unmixing algorithm always captures the entire duration of the root event for better recognition performance, the dictionary needs to contain the root-patterns starting at all possible times. Another reason for doing this is such that we obtain an overcomplete dictionary and the sparsity constraint would enable extraction of only a few large valued non-zero coefficients in vector α, reducing the false-alarm rate. In the following, we detail the construction process of the *temporal root-patterns*.

Suppose the root-patterns last for at most t_s seconds and the sampling frequency is ω, then each root-pattern would have $t_s \times \omega$ samples. For an observation vector $\mathbf{x}$ acquired from a FDR that lasts for t_c seconds, it can have $t_c \times \omega$ samples starting from the pre-event steady state, like 60 Hz, to the post-event steady state. The dimension of $t_c \times \omega$ is always larger than or equal to that of $t_s \times \omega$ since $t_c \geq t_s$ is always true, i.e., a multi-event observation always consists of at least one root event.

From the temporal perspective, each root-pattern can start at all possible sample points from the first one to the $(t_c - t_s) \times \omega$th during the time period t_c. To obtain

such temporal root-patterns, we construct $(t_c - t_s) \times \omega$ number of profiles for each root-pattern by shifting it from the first sample to the $(t_c - t_s) \times \omega$th sample. For instance, denote a root-pattern as $\mathcal{R}(k), k = 1, \ldots, t_s \times \omega$. If we want to construct one temporal root-pattern $\mathcal{T}(k), k = 1, \ldots, t_c \times \omega$ occurring at the ith sample point for $\mathcal{R}(k)$, then $\mathcal{T}(k)$ can be calculated as:

$$\mathcal{T}(k) = \begin{cases} \mathcal{R}(k) \otimes \delta(k - i) \text{ for } 1 \leq k < i + t_s \times \omega \\ \mathcal{R}(t_s \times \omega) \text{ for } i + t_s \times \omega \leq k \leq t_c \times \omega \end{cases} \tag{5.5}$$

where $\delta(k - i), k = 1, \ldots, t_c \times \omega$ is the Dirac δ function and $\otimes$ denotes convolution process. Since we only analyse frequency fluctuation of the signal for event unmixing purpose, the starting value of $\mathcal{R}(k)$ is 0Hz (by removing the base frequency 60Hz). Therefore, the function in Eq. (5.5) actually pads the samples before the selected ith point with zeros and the samples beyond $t_s \times \omega + i$ with the tail stable value of $\mathcal{R}(k)$ to form a temporal root-pattern. The whole formation of a group of temporal root-patterns derived from one root-pattern is illustrated in Figure 5.2.

5.2.3 Dictionary Augmentation

We denote all the "root-patterns" extracted from four event categories by $\mathbf{T}$ and all the "temporal root-patterns" via shifting and padding of $\mathbf{T}$ by $\widehat{\mathbf{T}}$, then the over-determined dictionary $\mathbf{S} = \widehat{\mathbf{T}}$. Ideally, the frequency signal of a multi-event that mixed by cascading constituent events can be precisely unmixed by the root-patterns in $\widehat{\mathbf{T}}$, producing correct number, type and occurring time of the constituent events. However, due to the strong correlation among devices in multi-event disturbance, the constituent event signals from the same device are not identical in different cases of multi-event. In addition, the transferable root-patterns are also not exactly the same as the constituent event signals of each device in multi-event. Therefore, due to this pattern variations, the unmixing result from Eq. (5.2) might pull in other temporal root-patterns that actually dose not happen to minimize the reconstruction error.

To solve this problem, we treat the pattern of each constituent event as a degraded version of the corresponding root-pattern in $\widehat{\mathbf{T}}$, and handle the distortion using a set of unit vectors, referred as "*trivial-patterns*" (Wright et al., 2009). Each trivial-pattern $\mathbf{g}_j$ has only one non-zero entry, 1, at jth position, $j = 1, 2, \ldots$. The trivial-patterns can be arranged in order to form an identity matrix $I = [\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_{l=t_c \times \omega}] \in \mathbb{R}^{l \times l}$. Then, in the unmixing step, the reconstruction fidelity and the sparsity constraints can still pick the correct root-patterns in $\widehat{\mathbf{T}}$ by activating some of the trivial-patterns for making up the error caused by the pattern variations, at the same time prevent the introduction of incorrect temporal root-patterns. In this way, the event detection, recognition and occurring time identification of each constituent event can still be accurate. In addition, we enforce nonnegativity constraint on the coefficients in α. To enable the trivial-patterns also make up negative reconstruction error, the "negative trivial-patterns" are also included. Consequently, the densely constructed dictionary $\mathbf{S}$ is now rewritten as below, and the corresponding unmixed coefficient vector α is composed by as in Eq. (5.6).

$$\mathbf{S} = [\widehat{\mathbf{T}},\ \mathbf{I},\ -\mathbf{I}], \qquad \alpha = [\alpha_{Eq.(5.2)},\ h^+,\ h^-] \tag{5.6}$$

5.2.4 Nonnegative Sparse Linear Unmixing

Once the over-determined dictionary $\mathbf{S}$ is constructed, we can apply it to estimate the coefficient vector α, which will be used for event detection – coefficients of non-zero value indicate that the corresponding temporal root-pattern in the dictionary should be used to reconstruct the observation mixture, $\mathbf{x}$. Since the temporal root-pattern indicates event's pattern occurred at a certain time, by deriving α, we can detect existence of multiple cascading events as well as their temporal locations (or starting time). Given the overcomplete dictionary, traditional methods for coefficient estimation such as fully constrained least squares (FCLS) (Heinz and Chein-I-Chang, 2001) or nonnegatively constrained least squares (NCLS) (Du et al., 2000) would

not work, as the estimated coefficient by FCLS and NCLS may have non-zero values on each root-pattern which would not serve as suitable ways for detection purposes. Recall that the number of constituent events in a multi-event is generally small.

Recent developments on sparse coding technique (Tibshirani, 1996; Candes and Tao, 2006; Lee et al., 2007) provides good solutions to the proposed NSEU algorithm in Eq. (5.2). The sparse coefficient vector produced by sparse coding has only a few entries being non-zero, indicating that only a few temporal root-patterns utilized to reconstruct the original signal $\mathbf{x}$. In the application of event detection, it is equivalent to say that only a few constituent events occurred. Although the sparse optimization problem in Eq. (5.2) is NP-hard in general, Donoho (Donoho and L., 2006) suggested that as long as the desired coefficient vector α is sufficiently sparse, it can be efficiently recovered by minimizing the ℓ^1-norm, and finally expressed as Eq. (5.8)

$$\min \|\alpha\|_1 \text{ s.t. } \|\mathbf{S}\alpha - \mathbf{x}\|_2^2 \leq \epsilon,\ \alpha \geq 0 \tag{5.7}$$

$$\alpha = \arg\min_{\alpha} \|\mathbf{x} - \mathbf{S}\alpha\|_2^2 + \lambda\|\alpha\|_1,\ \alpha \geq 0 \tag{5.8}$$

where λ balances the sparsity of the solution and the fidelity of the approximation to $\mathbf{x}$. The ℓ^1 norm is to enforce the sparsity on α. Eq. (5.8) is convex in α with $\mathbf{S}$ being fixed. The "feature-sign search algorithm" in (Lee et al., 2007) is revised to solve this sparse coding problem with the nonnegativity constraint added.

Event detection is mainly based on studying the non-zero coefficients in the sparse coefficient vector, α, as each coefficient corresponds to the weight of each atom (or the column in dictionary) in forming the observation vector $\mathbf{x}$. The larger the coefficient, the more contribution the root pattern has in making the mixture. An empirically determined threshold is applied on α, where coefficients above the threshold would correspond to the temporal root-patterns actually occurred at a certain time.

5.3 Experiments

5.3.1 Evaluation with Simulated Data

We conduct a series of experiments to demonstrate the effectiveness of the proposed NSEU approach using simulated data in this subsection. The simulations are done based on a small synthetic power grid model "savnw" using the software "Power System Simulator for Engineering (PSS/E) *". The grid model "savnw" is an example bench case supplied with PSS/E with configuration shown in Figure 5.3. The model includes 6 generators, 17 branch lines, 7 loads, and 21 buses. Each type of single event causes the system to arrive at a new steady state within 30 seconds. We thus use 30-second frequency fluctuation of single event for learning the root-patterns. Because this power grid model is quite small, we simply select 4 generator trips (GT) and 1 load shedding (LS) as root-patterns. As for the line trips (LT), we apply the K-means method to extract 5 root-patterns using the training data collected from 17 lines. Since we do not have oscillation data, we just omit the oscillation event for simplicity. The dictionary $\mathbf{S}$ is thus built based on the 10 extracted root-patterns. Suppose a multi-event case would last for 40 seconds and the sampling rate is set to 60Hz, the dictionary will then include $(40 - 30) \times 60\text{Hz} \times 10 = 6,000$ temporal root-patterns. We can improve the on-line performance by decreasing the number of temporal root-patterns in $\mathbf{S}$ by shifting at every two or more samples, at the sacrifice of losing certain temporal localization accuracy. In each test, the regularization parameter of NSEU is selected as $\lambda = 0.1$ and the threshold is selected as 0.034.

Detection Results

Single Event Detection and Recognition

We start with the detection of single event on some simulated cases. An example case of generator trip is discussed in this part. This single event was simulated for a

*PSSE, a power system simulation software provided by Siemens

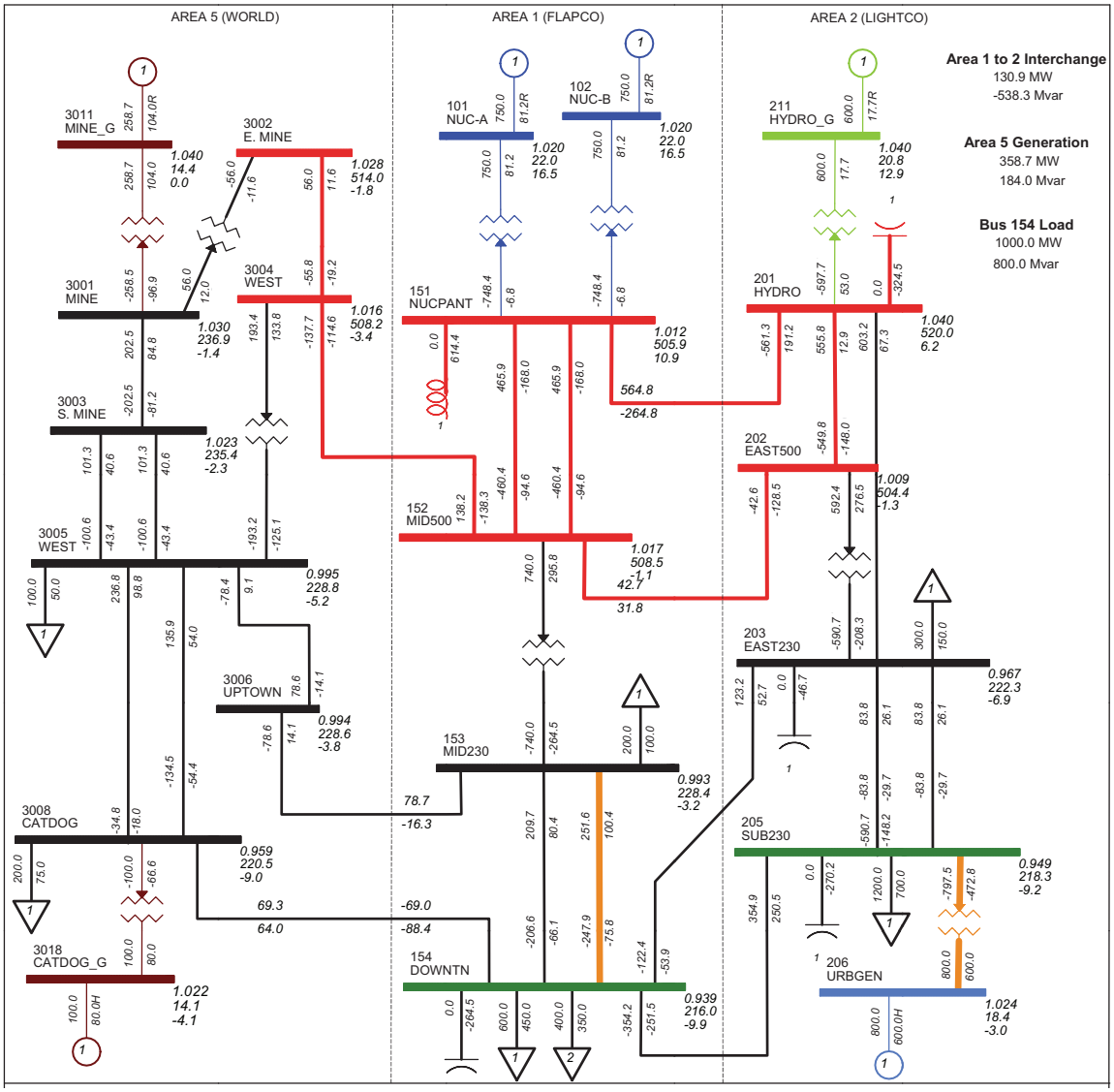

Figure 5.3: Configuration of the synthetic power grid model, "savnw", in PSS/E.

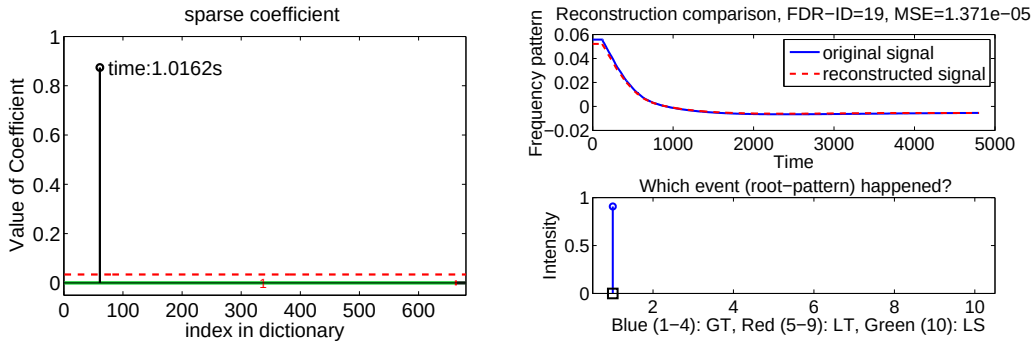

Figure 5.4: Simulated single event of GT101 at 1s, left: unmixed sparse coefficients α with event starting time, top-right: original and reconstructed signal, bottom-right: event type classification and ground truth marked with a black square.

duration of 40-sec and the trip occurred at the 1^{st} second. The frequency recording from a randomly selected bus is used to simulate data collected from a FDR. The unmixing result using the proposed NSEU are shown in Figure 5.4. The left sub-figure illustrates the estimated coefficient vector α, which is a sparse vector with only one non-zero coefficient, and the time index associated with this coefficient (i.e., the event starting time) is 1.02s, which is consistent with the ground truth that starts at the 1^{st} second. The top-right sub-figure demonstrates that the reconstructed frequency signal by $\mathbf{S}\alpha$ resembles the observed frequency signal, $\mathbf{x}$, from the simulated FDR. In the bottom-right sub-figure, all the coefficients belonging to the same root-pattern are summed up for recognition purpose, in order to identify what type of event has occurred. The indices for root-pattern $1 \sim 4$, $5 \sim 9$, 10 represent generator trip, line trip, and load shedding, respectively, and the index with a black square is the ground truth. This detection result indicates it is a generator trip of root-pattern 1, which agrees with the ground truth.

Multiple Event Detection and Recognition

Next, we show three example cases of simulated multi-event. Again, a randomly selected bus is used to extract the frequency signal representing data collected from a FDR. The first example consists of two cascading events: LT 201-202 and GT 3018 occurring at the 1^{st} and the 10^{th} second, respectively, recording frequency variations

for duration of 40s. The NSEU detection result is shown in Figure 5.5. The left sub-figure shows that there are several non-zero coefficients stay in two obvious compact clusters at time around 1.03s and 10.14s, based on which we can conclude that there are two events occurred around the 1^{st} and the 10^{th} second. The top-right sub-figure shows that the reconstructed signal is very close to the observed signal. In the bottom-right sub-figure, the approach determines that the two events are from root-patterns 7 and 4, which is consistent with the ground truth.

The second example also contains two cascading events: GT 101 and GT 3011 occurred at the 1^{st} and the 10^{th} second, respectively, recording frequency variations for a duration of 40s. Detection results shown in Figure 5.6 indicate that the approach detects two events from root-patterns 1 and 3 starting at 1.06s and 11s, respectively. It is a little bit imprecise for detection result of the second event GT 3011 occurring at 10^{th} second from root-pattern 4. This might due to the fact that 3011 is a swing generator with flexible power output to make the whole system to be resilient. GT 3011 happened after GT 101, thus its outputted power will increase a bit to make up the power lose in system after GT 101, therefore its frequency pattern will vary correspondingly, resulting in an inaccurate recognition result as from root-pattern 3 (but still GT). As for the 1s delay of the identified occurring time of GT 3011, we consider it is acceptable in heuristic. In addition, this error would not cause significant side-effect as the reason for deriving occurring time is to estimate the location of event where only the difference between the starting time sensed at different FDR sensors will be used for estimation. Fortunately, the number of swing generators in real power systems is very small and most of the generators are with fixed power output.

Finally, we show a simulated example of three cascading events: LT 154-3008, LT 151-201, and GT 3018 occurring at the 1^{st}, 8^{th}, and 15^{th} second, respectively, with a recording length of 50s. The detection result is shown in Figure 5.7. The reconstructed signal resembles the original input signal with just very small deviation. From the left sub-figure, we clearly see that there are three events above the detection threshold

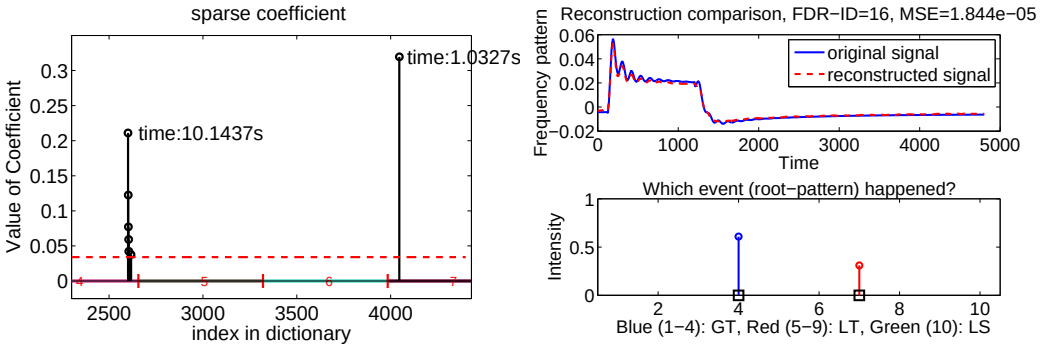

Figure 5.5: Simulated multi-event case of LT201-202 at 1s and GT3018 at 10s, left: unmixed sparse coefficients vector α with events starting time, top-right: original and reconstructed signal, bottom-right: event type classification indicates two events are from root-patterns 7&4, and ground truth is marked with black squares.

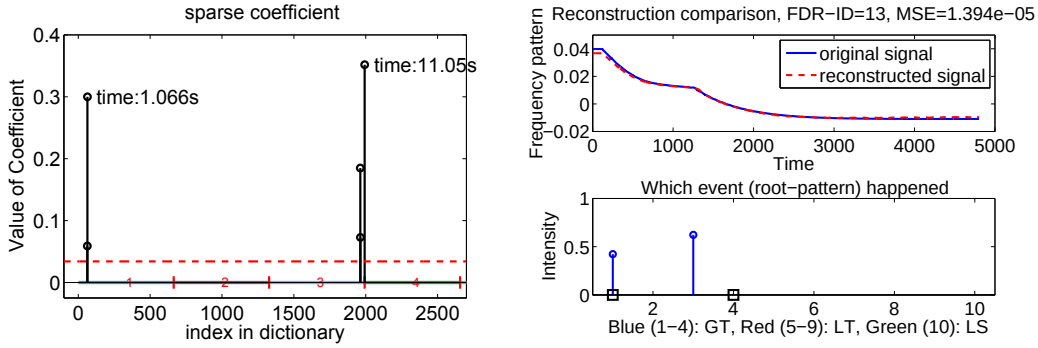

Figure 5.6: Simulated multi-event case of GT101 at 1s and GT3011 at 10s, left: unmixed sparse coefficients vector α with events starting time, top-right: original and reconstructed signal, bottom-right: event type classification indicates two events are from root-patterns 1&3, and ground truth is marked with black squares.

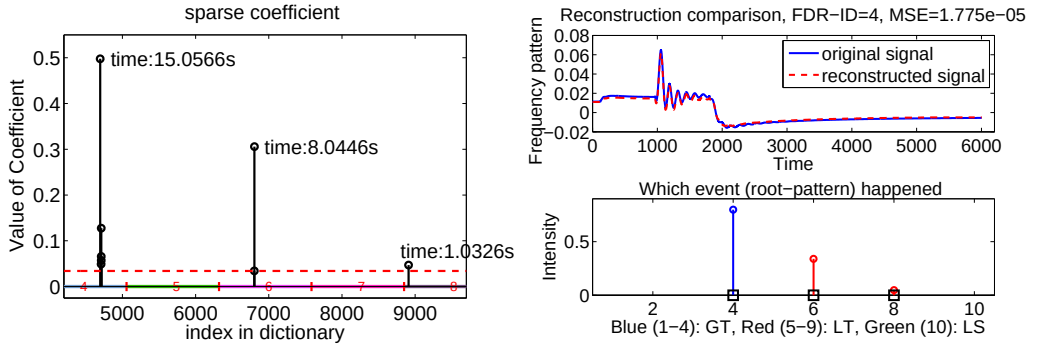

Figure 5.7: Simulated multi-event case of LT154-3008 at 1s, LT151-201 at 8s and GT3018 at 15s, left: coefficients vector α with events starting time, top-right: original and reconstructed signal, bottom-right: event type classification indicates three events are from root-patterns 8&6&4, and ground truth is marked with black squares.

Table 5.1: Quantitative evaluation on simulated event cases

	Detection	FalseAlarm	R-P Recog	OT-Deviation
S1C	100%	0%	100%	0.0579s
M2C	100%	5%	90%	0.1331s
M3C	100%	6.67%	86.7%	0.1056s

occurred around the 1.03s, 8.04s, and 15s, respectively. Combined with the bottom-right sub-figure, we can conclude that the first two events are line trips and the third is a generator trip, which are exactly what happened according to the ground truth.

Summary of Unmixing Performance

The above results suggest that the constraints in NSEU effectively confine the solution space, leading to accurate multiple events detection and recognition. We conducted experiments on 8 single event cases (S1C), 10 multi-event cases with two constituent components (M2C), 5 multi-event cases with three constituent components (M3C) in total. We use four metrics to measure performance, including detection accuracy, false alarm rate, root-pattern recognition rate (R-P Recog), and occurrence time deviation from the ground truth (OT-Deviation), as shown in Table 5.1.

Both detection accuracy and false alarm rate are used to evaluate the detection performance of NSEU approach. The detection accuracy calculates the ratio between the number of correctly detected and that of the total constituent events. The false alarm rate calculates the ratio between the number of detected but not really happen and that of the true constituent events according to the ground truth. As indicated in Table 5.1, we can find that the proposed NSEU approach detects all constituent events with 100% accuracy while with very low rate of false alarms.

The root-pattern recognition rate (R-P Recog) is used to evaluate the recognition performance. It calculates the ratio between the number of correctly identified events (i.e., events with correct type of root-pattern) and that of correctly detected events. Results in Table 5.1 indicate the averaged root-pattern recognition rates are pretty

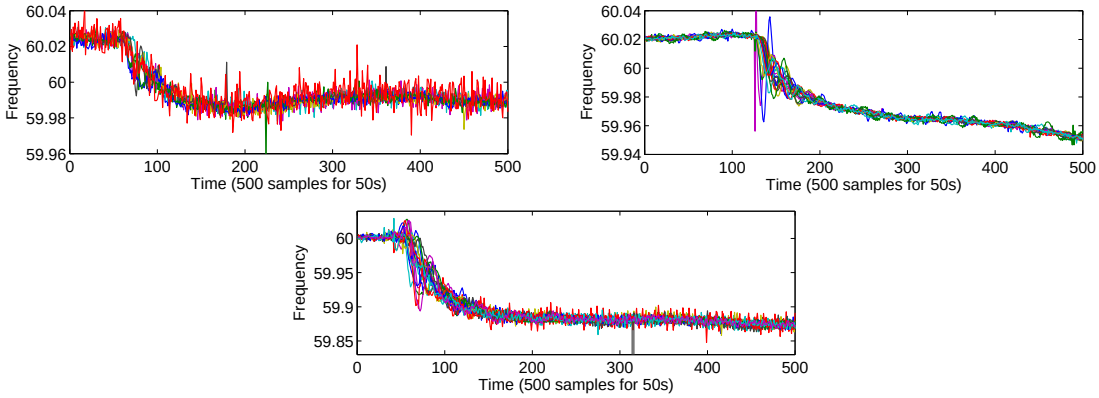

Figure 5.8: Frequency signals of three real event example cases, top, case 1: single event of a GT (10 FDR signals); middle, case 2: multi-events of one GT with one LT (18 FDR signals); bottom, case 3: multi-events of two GTs and two or three LTs (18 FDR signals).

high. The deviation between the identified occurring time (OT-Deviation) and the ground truth is used to evaluate the precision of temporal localization. As shown in Table 5.1, all the averaged OT-Deviation values are quite small, indicating the high accuracy of temporal localization.

5.3.2 Evaluation with Real Event Data

For evaluation with real event cases, we conducted similar experiments on real data. However, the root-patterns are learned from real data collected from both US-wide FNET (generator trip and load shedding events) and PSS/E (the line trip events).

Event Data and Preprocessing

Three real event cases are used for evaluation.

Case 1: A single event case (generator trip) happened at a power plant on May 30, 2006, as shown in Figure 5.8 (top). The first 50-second data (which is sampled at every 0.1s with a total of 500 samples) is used, and signals from 10 FDRs at different locations are used for checking the detection repeatability.

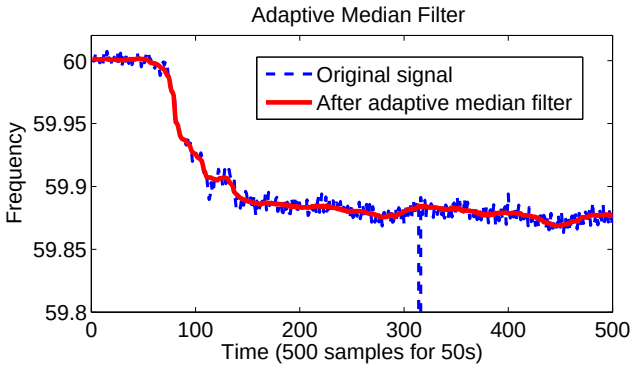

Figure 5.9: Applying the adaptive median filter on a signal collected from the FDR 14 of case 3. The filter successfully removed white noise and spikes, especially the large spike around the 32^{th} second.

Case 2: A multiple-event case (a generator trip followed by a line trip) happened at Surry, VA on February 2, 2011, as shown in Figure 5.8 (middle). We use the first 50-second data, and signals from 18 FDRs are used for checking the repeatability.

Case 3: A multiple-event case happened on August 4, 2007, comprised of multiple single-line-to-ground faults on a 765-kV line and generator trips at two locations. The Eastern Interconnection frequency thus dropped from 60 to 59.864 Hz, as shown in Figure 5.8 (bottom). The first 60-second data is used as most events occurred in this period (from stable 60Hz to the next stable 59.864Hz). Again, signals from 18 FDRs at different locations are used for checking the repeatability.

It is necessary to eliminate noises in the frequency data before applying the NSEU approach for event detection. In general, two kinds of noises contained in frequency signals, including white noise and impulsive noise. An adaptive median filter by (Li et al., 2010) is used for denoise. Figure 5.9 shows an example of the filtering effect. The filter successfully removed white noise and spikes in the original frequency signal, preventing possible detection error caused by undesired noises.

Root-Pattern Learning

Frequency data from individual generation trip and load shedding events are retrieved from FNET database (FnetDatabase, 2010). Since FNET does not detect line trips

Table 5.2: Breakdown of training event cases from Eastern (EI), Western (WECC) and Texas (ERCOT) interconnections.

	EI	WECC	ERCOT		EI
Generation Trip	547	415	189	Line Trip	257
Load Shedding	160	346	0		

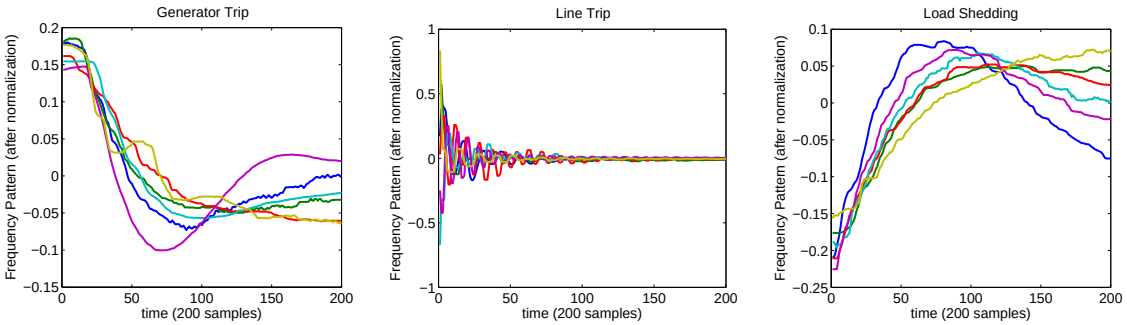

Figure 5.10: K-means clustering results for root-pattern learning (all the patterns above are normalized after remove their mean value).

currently, there is no entries or corresponding data for this event type. PSS/E was instead used to perform simulations of line trips. A 16,000-bus model of the Eastern Interconnection was used for the simulations. About 75 buses corresponding to actual FDRs were selected as measurement points, and lines adjacent to these buses were tripped one at a time. A 20-second simulation was performed in each case, with measurement points being saved at 0.1-second intervals to match the rate of FDRs. Finally, all the data selected from different sources for training is shown in Table 5.2.

K-means: K-means clustering is used for root-pattern learning. We set $K = 6$ and 6 root-patterns are learned for each of the three event categories. As shown in Figure 5.10, we confirm that the learned root-patterns of each category indeed share similar characteristics with certain degree of difference, as explained in Sec. 5.2.2.

Detection Results

The root-patterns learned by K-means are used to construct temporal root-patterns, which form the dictionary $\mathbf{S}$. The detection threshold is set as 0.04.

Table 5.3: Event detection results for case 1, one generator trip actually happened in this real event, and all the FDRs successfully detected the generator trip.

FDR	1	2	3	4	5
GenTrip1	8.4s	8.4s	8.8s	8.4s	8.6s
FDR	6	7	8	9	10
GenTrip1	9.0s	7.8s	7.2s	8.4s	7.8s

Table 5.4: Event detection results for case 2, where one generator trip and one line trip actually happened in this real event. All the FDRs successfully detected one generator trip (root-pattern 6) and one line trip (root-pattern 8 or 12).

FDR	1	2	3	4	5	6	7	8	9
GenTrip6	12.4s	14.8s	12.6s	13.6s	12.2s	12.4s	12.4s	11.6s	12.2s
LineTrip8			14.2s		15.2s	14.2s	14.2s	13.6s	
LineTrip12	13.8s	14.2s		13.8s					15.2s
FDR	10	11	12	13	14	15	16	17	18
GenTrip6	12.4s	14.6s	14.2s	12.6s	14.2s	11.8s	13.6s	12.2s	12.4s
LineTrip8				15.2s				15.6s	15.8s
LineTrip12	14.2s	13.4s	13.8s		13.8s	15.6s	13.8s		

Case 1: case 1 is a single event case. The detection results, as shown in Table 5.3, indicate that the NSEU approach successfully detected a single generator trip of root-pattern 1 from each signal of 10 FDRs at different locations without any false alarm. An example of the detected temporal root-patterns and the reconstructed signal from FDR 2 are shown in Figure 5.11. We can find from the top-right sub-figure that the reconstructed signal closely resembles the original signal. In the left sub-figure, there are several large coefficients appear overlapped in the first root-pattern area. These lines in fact belong to the same event, as they are quite temporally close, indicating one generator trip as the only major event at 8.4s. The other coefficients smaller than threshold are only useful to minimize the reconstruction error.

Case 2: case 2 is a multi-event case with one generator trip followed by one line trip. The detection results are shown in Table 5.4. All the signals from 18 FDRs are successfully detected one generator trip from root-pattern 6 and one line trip, though the detected line trip may be from either root-pattern 8 or 12. This may because the line trip's root-patterns learned based on PSS/E simulations can not reflect very well

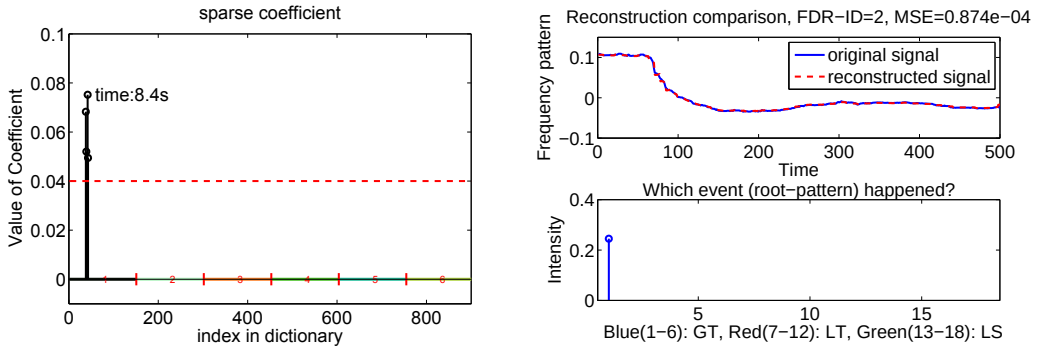

Figure 5.11: Case 1 detection result using data from FDR 2, one generator trip is detected at 8.4s. Left: coefficients of the detected root-pattern; Top-right: original event signal and reconstructed signal; Bottom-right: the detected event is a GT from the first root-pattern out of 18.

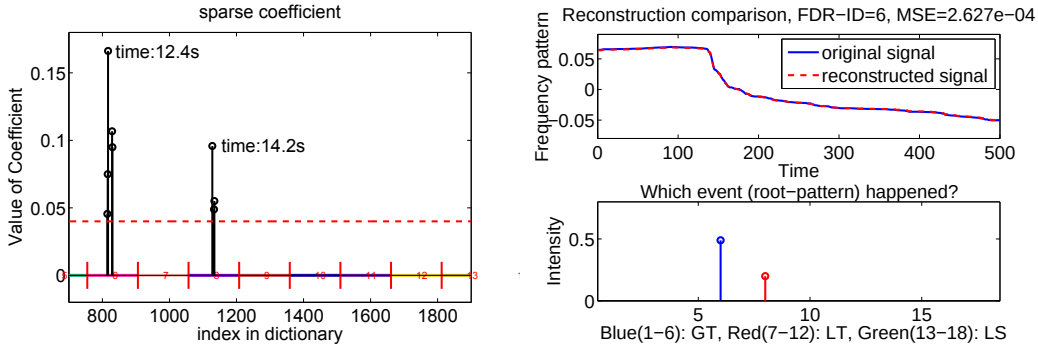

Figure 5.12: Case 2 detection result using data from FDR 6, one generator trip and one line trip are detected at 12.4s and 14.2s, respectively. Left: coefficients of the detected root-patterns; Bottom-right: two detected events including one GT and one LT from the sixth and the eighth root-patterns.

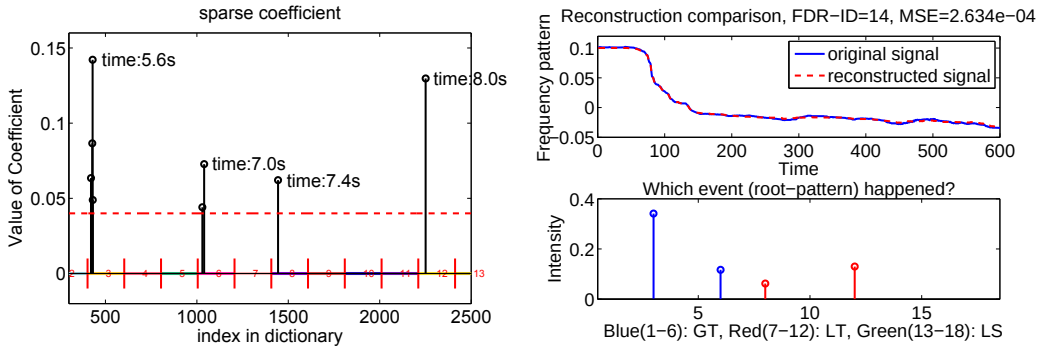

Figure 5.13: Case 3 detection result using data from FDR 14, two generator trips are detected at 5.6s and 7.0s, and two line trips are detected at 7.4s and 8.0s, respectively. Bottom-right: four detected events include two GTs and two LTs from the third, sixth, eighth and twelfth root-pattern, respectively.

Table 5.5: Event detection results for case 3, two generator trips and multiple line trips might have occurred in this real event. Most FDR signals detected two generator trips (root-patterns 3&6) and two line trips (root-patterns 8&12), but the generator trip root-pattern 3 was not detected by FDR 2&16 and the line trip root-pattern 12 was not detected by FDR 3.

FDR	1	2	3	4	5	6	7	8	9
GenTrip3	5.4s		4.0s	7.0s	5.2s	4.6s	4.4s	4.0s	4.6s
GenTrip6	6.2s	4.2s	3.2s	7.6s	4.6s	4.2s	6.4s	3.8s	6.8s
LineTrip8	9.2s	7.2s	7.8s	7.6s	9.0s	6.2s	6.8s	5.6s	8.6s
LineTrip12	7.2s	6.2s		5.6s	6.4s	7.4s	7.2s	5.8s	7.2s
FDR	10	11	12	13	14	15	16	17	18
GenTrip3	4.2s	4.0s	4.2s	5.0s	5.6s	4.2s		3.6s	4.4s
GenTrip6	4.0s	6.0s	5.4s	4.2s	7.0s	3.8s	4.4s	3.2s	7.8s
LineTrip8	6.0s	6.6s	8.6s	9.0s	7.4s	6.8s	7.6s	7.8s	6.2s
LineTrip12	6.0s	6.6s	7.2s	7.0s	8.0s	7.2s	7.4s	5.8s	6.4s

the behavior of real line trip event. An example of the detected temporal root-pattern and the reconstruct signal from FDR 6 are shown in Figure 5.12. The top-right sub-figure shows that the reconstructed signal is quite close to the original signal. In the left sub-figure, the unmixed sparse coefficients are closely clustered in root-patterns 6 and 8, indicating the NSEU is able to correctly unmix this multi-event as a generator trip followed by a line trip.

Case 3: case 3 is another multi-event case with two generator trips and possibly two or three line trips involved, thus it is more complicated than the cases 1 and 2. The detection results are shown in Table 5.5, which demonstrate the NSEU approach successfully detected two generator trips from 16 out of 18 FDRs and two line trips from 17 FDRs without false alarm. An example of the detected root-patterns from FDR 14 is shown in Figure 5.13. From the left sub-figure of the sparse coefficients, it is clear that the approach detects two generator trips and two line trips correctly. The reconstructed signal in top-right sub-figure is also very close to the observed signal. This results further confirm the effectiveness of NSEU approach, however, line trips in this case are not completely detected as one line trip is missed. This might due to the too trivial frequency change brought by the miss-detected line trip.

Table 5.6: Quantitative evaluation on real event cases (FA: false alarm ratio).

	FDR Num	GT Det	GT FA	LT Det	LT FA
Case1	10	100%	0%		
Case2	18	100%	0%	100%	0%
Case3	18	94.4%	0%	64.8%	0%
Mean		98.15%	0%	82.4%	0%

Summary of Unmixing Performance

The three experiments with real event cases analysed frequency signals from 46 FDRs in total. As shown in Table 5.6, the proposed NSEU approach detects constituent events with 98.15% averaged accuracy for generator trip and 82.4% averaged accuracy for line trip without false alarm. Due to the lack of ground truth, we cannot evaluate performance of recognition or temporal localization.

The experimental results demonstrate advantages of the proposed NSEU approach over the other existing event detection techniques. In Figure 5.8, we can observe that there is no immediately perceivable difference between frequency signals of the multi-event case and that of the single-event case, because the mixing process will occlude or degrade most of the features from different root-patterns. Most existing techniques based on immediately detectable information can only detect the starting time of the initial event involved in multi-cascading-event. In contrast, the NSEU approach is able to uncover the constituent root events with high detection accuracy. In addition, the difficulty for line trips detection can be attributed to 3 aspects: First, the root-patterns of line trip are learned via simulations, which may not reflect the dynamics of line trips that occurred in real world. Second, the frequency change caused by line trips is generally smaller than that of generator trips, probably causing the unmixed coefficients on line trips to be much smaller than that on generator trips. Therefore, line trips may not be detected if using the same detection threshold for GTs and LSs. Third, the power imbalance caused by some line trips is quite tiny that can be easily adjusted by system's self-resilience and thus not perceivable.

5.4 Summary

This chapter presented a novel interpretation of systematics of the frequency signal formation of the multi-events in smart power grid. Through analysis of the connection between frequency disturbance caused by multi-events and that by single-events, we extracted a set of transferable root-patterns and developed an effective and promising constrained "event unmixing" approach, NSEU, based on a linear mixture model for constituent events detection, recognition, and temporal localization in disturbance of a multi-cascading-event. The experimental results with both simulated and real event data demonstrated the effectiveness of the proposed approach.

Benefited by the frequency data recorded by FNET, this work provides a new and feasible way to obtain high-resolution situational awareness for smart grid system. The findings in this work would also benefit other real applications, particularly for example, microgrids remote monitoring, multiple events localization, and smart grid coordination.

Chapter 6

Conclusion and Future Work

This chapter summarizes the study observations and discusses possible improvements for future research.

6.1 Summary

Traditional machine learning approaches assumes the training and testing samples are from the same domain, thus the model learned from the training samples can be adapted to the testing samples directly. However, most of the practical applications cannot guarantee this assumption. Due to various factors in data collection, such as different recording viewpoints, various time of data sampling, diverse environments, the data collected for model learning dose not always have the same distribution as that in testing. To precisely and robustly recognize targets across different data domains is a fairly challenging problem, the key issue for the solution is to find out the latent relationship between the different data domains, and build an effective connection to facilitate the real world recognition problems. In this dissertation, how to exploit the latent relationships across different data domains was studied. We focused on three typical but inner related cross domain recognition problems, from 3-D video to 1-D signal, i.e., (1) action recognition across camera views, (2) person

re-identification across camera views, (3) multi-event detection and recognition in smart grid system.

First, to solve the problem of action recognition across camera views, we make use of learning samples from different camera domains to learn a reconstructable path between any two camera views. The reconstructable path is able to exploit structure information in each view domain and well preserve the category discrimination. In addition, the seemingly useless samples are also made use of to improve the learning performance. Extensive experimental results show that our approach achieves very competitive performance compared to the state-of-the-art approaches. Second, to solve the problem of person re-identification in non-overlapped camera networks, we make use of the paired training samples from two camera domains to learn locally constrained adaptive distance metrics based on a random kernel forest. Our approach discriminatively assign each local patch of a query image to the optimal local kernel for distance measure, therefore the distance between images from the same individual can be well minimized than that from two different people. Again, experiments show the effectiveness of our approach compared to the state-of-the-art. Finally, to solve the problem of multi-event detection and recognition in smart grid, we supposed the root patterns that embedded in the single events domain and multi-events domain are similar and thus can be transferable used. Based on this assumption, we proposed the NSEU algorithm to achieve the simultaneous detection and recognition of each constituent component event using frequency signals. Up to our best knowledge, this is the first realization of multi-event analysis with high accuracy in smart grid.

6.2 Future Research

Our future research lies in two main aspects.

- First, our approach for action recognition still requires learning samples, either unlabelled or semi-labelled data, from both of the two cameras for extracting

relationships between the two view domains. In the future work, we will try to find the third party resources for building the inter-between connections, e.g., 3D skeleton data, which can be adapted to any camera view. Furthermore, the latent structure of the action motion information should be view-invariant, how to extract the motion structure is also promising to realize the practical view-invariant action recognition problem. Deep learning is demonstrated the best tool for feature extraction and widely applied in various computer vision tasks. However, how to make use of deep learning to extract the temporal structure of motion information that is unique for video based action recognition still has not been touched much. Recently, recurrent neural network (RNN) is designed to mine the temporal information in video analysis, we will follow this cutting-edge technique and put our effort on applying RNN onto our cross view action recognition task.

- Second, our approach for person re-identification learned a random kernel forest that is able to assign different local region a specific but also optimal local metric kernel, which enables the images from the same individual to have the minimal distance. However, our local distance metrics only consider how to minimize the pairwise distance between true image pairs recorded from disjoint cameras. The re-identification performance can surely be improved by mining the hard negatives, i.e., how to magnify the distance between false image pairs. To learn such a more discriminative *local distance metric*, we also propose to find a local projection $\mathbf{p}_k$ within each local transform that maximize the objective function:

$$\mathbf{p}_k^* = \arg\max_{\mathbf{p}_k} \frac{\sum_{i \in G_k} ||\mathbf{p}_k x_i - \mathbf{p}_k y_i^-||^2}{\sum_{i \in G_k} ||\mathbf{p}_k x_i - \mathbf{p}_k y_i^+||^2} \tag{6.1}$$

  Now each entry in the group of learning samples for each local kernel becomes a triplet $\{(x_i, y_i^+, y_i^-)_{i=1,\ldots}^n\}_{i,n \in G_k}$. This objective function has similar formulation as the linear discriminative analysis, we can also use Lagrange to convert it into a constrained optimization problem similarly.

Bibliography

Aharon, M., Elad, M., and Bruckstein, A. (2006). K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. on Signal Processing*, 54(11):4311–4322. 13

Ahmed, E., Jones, M., and Marks, T. (2015). An improved deep learning architecture for person re-identification. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 16, 52, 53

Arnold, A., Nallapati, R., and Cohen, W. W. (2007). A comparative study of methods for transductive transfer learning. In *IEEE Int. Conf. on Data Mining Workshop*. 3

Arnold, A., Nallapati, R., and Cohen, W. W. (2008). Exploiting feature hierarchy for transfer learning in named entity recognition. In *ACL:HLT*. 10

Baktashmotlagh, M., Harandi, M., Lovell, B., and Salzmann, M. (2013). Unsupervised domain adaptation by domain invariant projection. In *IEEE Int. Conf. on Computer Vision (ICCV)*. 10

Bazzani, L., Cristani, M., Perina, A., and Murino, V. (2012). Multiple-shot person re-identification by chromatic and epitomic analyses. *Pattern Recognition Letter*, 33(7):898–903. 14

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202. 13

Bergamo, A. and Torres, L. (2010). Exploiting weakly-labeled web images to improve object classification. In *IEEE Conf. on Neural Information Processing Systems (NIPS)*. x, xiii, 43, 44, 45

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*. Springer. 1

Blitzer, J., McDonald, R., and Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*. 3

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32. 16

Bruzzone, L. and Marconcini, M. (2013). Domain adaptation problems: adasvm classification technique and a circular validation strategy. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(2):770787. 4

Bulo, S. R. and Kontschieder, P. (2014). Neural decision forests for semantic image labelling. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 17

Bykhovsky, A. and Chow, J. (2003). Power system disturbance identification from recorded dynamic data at the northfield substation. *Int. Journal Electrical Power and Energy Systems*, 25(1):787–795. 78

Candes, E. and Tao, T. (2006). Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. on Information Theory*, 52(12):5406–5425. 83

Caruana, R. (1997). Multitask learning. *Machine Learning*, 28(1):41–65. 9

Chen, D., Yuan, Z., Hua, G., Zheng, N., and Wang, J. (2015a). Similarity learning on an explicit polynomial kernel feature map for person re-identification. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 15, 52

Chen, J., Zhang, Z., and Wang, Y. (2015b). Relevance metric learning for person re-identification by exploiting listwise similarities. *IEEE Trans. on Image Processing*, 24(12):4741–4756. 67, 70

Chen, L., Zhang, Q., and Li, B. (2014). Predicting multiple attributes via relative multi-task learning. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 10

Chen, S. S., Donoho, D. L., and Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1):33–61. 13

Cheng, D., Cristani, M., Stoppa, M., Bazzani, L., and Murino, V. (2011). Custom pictorial structures for re-identification. In *British Machine Vision Conference (BMVC)*. 14, 62

Cheng, L. and Pan, S. J. (2014). Semi-supervised domain adaptation on manifolds. *IEEE Trans. on Neural Networks and Learning Systems*, 25(12):2240–2249. 10

Chow, J., Vanfretti, L., Armenia, A., Ghiocel, S., and et al (2009). Preliminary synchronized phasor data analysis of disturbance events in the us eastern interconnection. In *IEEE PES Power Systems Conference and Exposition*. 18

Cong, Y., Yuan, J., and Liu, J. (2011). Sparse reconstruction cost for abnormal event detection. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 14

Criminisi, A. and Shotton, J. (2013). Decision forests for computer vision and medical image analysis. *Springer*. 54, 57, 58

Criminisi, A., Shotton, J., and Konukoglu, E. (2011). Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning. *Microsoft Research technical Report*. xii, 17

Dantone, M., Gall, J., Leistner, C., and Gool, L. V. (2013). Human pose estimation using body parts dependent joint regressors. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 17

Daume, H. and Marcu, D. (2006). Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26(1):101–126. 9

Ding, Z., Suh, S., Han, J.-J., Choi, C., and Fu, Y. (2015). Discriminative low-rank metric learning for face recognition. In *IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*. 10

Dollar, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). Behavior recognition via sparse spatiotemporal features. In *IEEE Int. Conf. on Computer Vision (ICCV) workshop VS-PETS*. 22, 37

Dollar, P. and Zitnick, C. L. (2013). Structured forests for fast edge detection. In *IEEE Int. Conf. on Computer Vision (ICCV)*. 17, 58

Dong, C., Zhao, H., and Wang, W. (2009). Hyperspectral image anomaly detection based on local orthogonal subspace projection. *Optics and Precision Engineering*, 17(8):2004–2010. 20

Dong, J., Zuo, J., Wang, L., Kook, K., Chung, Y., Liu, Y., and et al (2007). Analysis of power system disturbances based on wide-area frequency measurements. In *IEEE Power and Energy Society General Meeting (PESGM)*. 78

Donoho and L., D. (2006). For most large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829. 83

Du, Q., Chang, C.-I., Heinz, D., Althouse, M., and Ginsberg, I. (2000). A linear mixture analysis-based compression for hyperspectral image analysis. In *IEEE International Geoscience and Remote Sensing Symposium*. 82

Duan, L., Xu, D., Tsang, I. W., and Luo, J. (2011). Visual event recognition in videos by learning from web data. *IEEE Trans. on Pattern Analysis and Machine Intelligence*. 2

Elli, H. C. (1965). The transfer of learning. *The Macmillan Company*. 9

Fanello, S. R., Keskin, C., Kohli, P., Izadi, S., Shotton, J., Criminisi, A., Pattacini, U., and Paek, T. (2014). Filter forests for learning data-dependent convolutional kernels. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 17

Farenzena, M., Bazzani, L., Perina, A., Murino, V., and Cristani, M. (2010). Person re-identification by symmetry-driven accumulation of local features. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 14, 67, 69, 70

Farhadi, A. and Tabrizi, M. K. (2008). Learning to recognize activities from the wrong view point. In *European Conf. on Computer Vision (ECCV)*. 11, 12, 23

Farhadi, A., Tabrizi, M. K., Endres, I., and Forsyth, D. (2009). A latent model of discriminative aspect. In *IEEE Int. Conf. on Computer Vision (ICCV)*. x, 12, 41

Fernando, B., Habrard, A., Sebban, M., and Tuytelaars, T. (2013). Unsupervised visual domain adaptation using subspace alignment. In *IEEE Int. Conf. on Computer Vision (ICCV)*. 3

FnetDatabase (2010). FNET Event Database Search. *http://powerit.utk.edu/search*. 91

Friedman, J., Hastie, T., and Tibshira, R. (2010). A note on the group lasso and a sparse group lasso. *Technical report*. 29

Gardener, R. and Liu, Y. (2007). FNET: a quickly deployable and economic system to monitor the electric grid. In *IEEE Conf. on Technologies for Homeland Security*. 18, 73

Gardner, R., Wang, J., and Liu, Y. (2006). Power system event location analysis using wide-area measurements. In *IEEE Power Engineering Society General Meeting (PESGM)*. 18, 73

Gopalan, R. (2013). Learning cross-domain information transfer for location recognition and clustering. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 3

Gray, D. and Tao, H. (2008). Viewpoint invariant pedestrian recognition with an emsemble of local features. In *European Conf. on Computer Vision (ECCV)*. 15, 52, 62

Guillaumin, M., Verbeek, J., and Schmid, C. (2009). Is that you? metric learning approaches for face identification. In *IEEE Int. Conf. on Computer Vision (ICCV)*. 69, 70

Guo, R. and Qi, H. (2015). Facial feature parsing and landmark detection via low-rank matrix decomposition. In *IEEE Int. Conf. on Image Processing (ICIP)*. 20

Guo, R., Wang, W., and Qi, H. (2015). Hyperspectral image unmixing using cascaded autoencoder. In *IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (Whispers)*. 19

Hallman, S. and Fowlkes, C. C. (2015). Oriented edge forests for boundary detection. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 17

Harpale, A. and Yang, Y. (2010). Active learning for multi-task adaptive filtering. In *IEEE Int. Conf. on Machine Learning*. 4

Hastie, T., Tibshirani, R., and Friedma, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer. 2

Heinz, D. and Chein-I-Chang (2001). Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 39(3):529–545. 82

Hiraoka, Y., Shimi, T., and Haraguchi, T. (2002). Multispectral imaging fluorescence microscopy for living cells. *Cell Structure and Function*, 27:367–374. 20

Hirzer, M., Roth, P., Kostinger, M., and Bischof, H. (2012). Relaxed pairwise learned metric for person re-identification. In *European Conf. on Computer Vision (ECCV)*. 15, 67, 69, 70

Huang, C., Yeh, Y., and Wang, Y. F. (2012). Recognizing actions across cameras by exploring the correlated subspace. In *European Conf. on Computer Vision (ECCV)*. 12, 37

J. Chen, Z. Z. and Wang, Y. (2014). Relevance metric learning for person re-identification by exploiting global similarities. In *Int. Conf. on Pattern Recognition*. 69, 70

Jhuo, I.-H., Liu, D., Lee, D. T., and Chan, S.-F. (2012). Robust visual domain adaptation with low-rank reconstruction. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 4, 11

Ji, Y., Lin, T., and Zha, H. (2009). Mahalanobis distance based non-negative sparse representation for face recognition. In *IEEE Int. Conf. on Machine Learning and Applications*. 79

Jiang, Z., Lin, Z., and Davis, L. (2013). Label consistent k-svd: Learning a discriminative dictionary for recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(11):2651–2664. 13

Jing, X., Zhu, X., Wu, F., You, X., Liu, Q., Yue, D., Hu, R., and Xu, B. (2015). Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 15

Juneio, I. N., Dexter, E., Laptev, I., and Perez, P. (2008). Cross-view action recognition from temporal self-similarities. In *European Conf. on Computer Vision (ECCV)*. 10, 11, 23

Khorsandi, R., andM. Mottaleb, A. T., and Qi, H. (2015). Joint weighted dictionary learning and classifier training for robust biometric recognition. In *IEEE Global Conf. on Signal and Information Processing (GlobalSIP)*. 14

Kook, K. and Liu, Y. (2011). Wide-area frequency-based tripped generator locating method for interconnected power system. *Journal of Electrical Engineering and Technology*, 6(6):776–785. 18, 73

Kostinger, M., Hirzer, M., Wohlhart, P., Roth, P., and Bischof, H. (2012). Large scale metric learning from equivalence constraints. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 15, 69, 70

Krupka, E., Vinnikov, A., Klein, B., Hillel, A. B., and Freedman, D. (2014). Discriminative ferns ensemble for hand pose recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 17

Kulis, B., Saenko, K., and Darrell, T. (2011). What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 10

Layne, R., Hospedales, T., Gong, S., and Mary, Q. (2012). Person reidentification by attributes. In *British Machine Vision Conference (BMVC)*. 16

Lee, H., Battle, A., Raina, R., and Y. Ng, A. (2007). Efficient sparse coding algorithms. In *IEEE Conf. on Neural Information Processing Systems (NIPS)*. 13, 83

Lewandowski, M., Makris, D., and Nebel, J. (2010). View and style-independent action manifolds for human activity recognition. In *European Conf. on Computer Vision (ECCV)*. 23

Li, B., Camps, O. I., and Sznaier, M. (2012). Cross-view activity recognition using hankelets. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 12

Li, F. and Perona, P. (2005). A bayesian heirarcical model for learning natural scene categories. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 23, 37

Li, L., Li, S., and Fu, Y. (2013a). Discriminative dictionary learning with low-rank regularization for face recognition. In *IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*. 13

Li, R., Tian, T., and Sclaroff, S. (2007). Simultaneous learning of nonlinear manifold and dynamical models for high-dimensional time series. In *IEEE Int. Conf. on Computer Vision (ICCV)*. 11, 23

Li, R. and Zickler, T. (2012). Discriminative virtual views for cross-view action recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. x, xiii, 12, 25, 37, 38, 41, 42, 43, 44, 45

Li, S., Wang, W., Qi, H., Ayhan, B., Kwan, C., and Vance, S. (2015a). Low-rank tensor decomposition based anomaly detection for hyperspectral imagery. In *IEEE Int. Conf. on Image Processing (ICIP)*. 20

Li, T., Ding, C., Zhang, Y., and Shao, B. (2009a). Knowledge transformation for crossdomain sentiment classification. In *ACM SIGIR Conf. on Research and Development in Information Retrieval*. 3

Li, W., Duan, L., Xu, D., and Tsang, I. (2014a). Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 36(6):1134–1148. 10

Li, W., Tang, J., Ma, J., and Liu, Y. (2010). Online detection of start time and location for hypocenter in north american power grid. *IEEE Trans. on Smart Grid*, 1(3):253–260. 18, 19, 73, 91

Li, W. and Wang, X. (2013). Locally aligned feature transforms across views. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 61, 66, 67, 70

Li, W., Zhao, R., Xiao, T., and Wang, X. (2014b). Deepreid: Deep filter pairing neural network for person re-idnetification. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 16, 52, 53, 69, 70

Li, Y., Wu, Z., and Radke, R. (2015b). Multi-shot re-identification with random projection based random forests. In *IEEE Winter Conf. on Applications of Computer Vision*. 15

Li, Y., Zhou, Y., Xu, L., and Yang, X. (2009b). Incremental sparse saliency detection. In *IEEE Int. Conf. on Image Processing (ICIP)*. 14

Li, Z., Chang, S., Huang, F. L. T., Cao, L., and Smith, J. (2013b). Learning locally adaptive decision functions for person verification. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 15, 52, 67, 69

Liao, S., Hu, Y., Zhu, X., and Li, S. Z. (2015). Person re-identification by local maximal occurrence representation and metric learning. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 15, 69, 70

Lin, Z., Jiang, Z., and Davis, L. (2009). Recognizing actions by shapemotion prototype trees. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 22

Lisanti, G., Masi, I., Bagdanov, A. D., and Bimbo, A. D. (2015). Person re-identification by iterative re-weighted sparse ranking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 37(8):1629–1643. 67, 69

Liu, C., Gong, S., Loy, C., and Lin, X. (2012a). Person re-identification: what features are important? In *European Conf. on Computer Vision (ECCV)*. 14

Liu, J., Shah, M., Kuipers, B., and Savarese, S. (2011a). Cross-view action recognition via view knowledge transfer. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. x, 11, 12, 23, 37, 38, 41, 45

Liu, X., Song, M., Tao, D., Zhou, X., Chen, C., and Bu, J. (2014). Semi-supervised coupled dictionary learning for person re-identification. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 15, 61, 67, 69

Liu, X., Song, M., Zhao, Q., Tao, D., Chen, C., and Bu, J. (2012b). Attributerestricted latent topic model for person re-identification. *Pattern Recognition*, 11(2):334–345. 16

Liu, Y. (2006). A US-wide power systems frequency monitoring network. In *IEEE PES Power Systems Conference and Exposition*. 18, 73

Liu, Y., Xu, D., Tsang, I., and Luo, J. (2011b). Textual query of personal photos facilitated by large-scale web data. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(2):1022–1036. 3

Loy, C., Liu, C., and Gong, S. (2013). Person re-identification by manifold ranking. In *IEEE Int. Conf. on Image Processing (ICIP)*. 69, 70

Loy, C. and Tang, X. (2009). Multi-camera activity correlation analysis. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 51, 62

Luo, J. and Qi, H. (2010). Distributed object recognition via feature unmixing. In *ACM/IEEE Int. Conf. on Distributed Smart Cameras (ICDSC)*. 20

Luo, J. and Qi, H. (2012). Motion local ternary pattern for distributed human action recognition. In *ACM/IEEE Int. Conf. on Distributed Smart Cameras*. 22

Luo, J., Wang, G., Qi, H., Yokoyama, Y., Liaw, P. K., and Inoue, A. (2012). Interpreting temperature evolution of a bulk metallic glass during cyclic loading through spatial-temporal modeling. *Intermaterllics*, 29:1–13. 20

Luo, J., Wang, W., and Qi, H. (2013a). Feature extraction and representation for distributed multi-view human action recognition. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 3(2):145–154. 11

Luo, J., Wang, W., and Qi, H. (2013b). Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In *IEEE Int. Conf. on Computer Vision (ICCV)*. 14, 22

Luo, J., Wang, W., and Qi, H. (2014a). Spatio-temporal feature extraction and representation for rgb-d human action recognition. *Pattern Recognition Letter*, 50(1):139–148. 14

Luo, Y., Liu, T., Tao, D., and Xu, C. (2014b). Decomposition-based transfer distance metric learning for image classification. *IEEE Trans. on Image Processing*, 23(9):3789–3801. 10

Ma, A., Li, J., Yuen, P., and Li, P. (2015). Cross-domain person reidentification using domain adaptation ranking svms. *IEEE Trans. on Image Processing*, 24(5):1599–1613. 10

Ma, B., Su, Y., and Jurie, F. (2012a). Bicov: a novel image representation for person re-identification and face verification. In *British Machine Vision Conference (BMVC)*. 14

Ma, B., Su, Y., and Jurie, F. (2012b). Local descriptors encoded by fisher vectors for person re-identification. In *European Conf. on Computer Vision (ECCV)*. 15, 69

Ma, L., Yang, X., and Tao, D. (2014). Person re-identification over camera networks using multi-task distance metric learning. *IEEE Trans. on Image Processing*, 23(8):3656–3670. 15, 67, 70

Mairal, J., Bach, F., Ponce, J., and Sapiro, G. (2009). online dictionary learning for sparse coding. In *IEEE Int. Conf. on Machine Learning*. 13, 27, 29

Maji, S., Berg, A. C., and Malik, J. (2008). Classification using intersection kernel support vector machines is efficient. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 35

Markham, P. and Liu, Y. (2011). Artificial neural network-based classifier for power system events. In *Technical Report, University of Tennessee*. 78

Mcfee, B. and Lanckriet, G. (2010). Metric learning to rank. In *IEEE Int. Conf. on Machine Learning*. 70

Mehrotra, R., Agrawal, R., and Haider, S. A. (2012). Dictionary based sparse representation for domain adaptation. In *ACM Int. Conf. on Information and Knowledge Management*. 14

Mignon, A. and Jurie, F. (2012). Pcca: A new approach for distance learning from sparse pairwise constraints. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 15, 67, 69, 70

Naqvi, S., Khan, M., Barnard, M., and Chambers, J. (2012). Multimodal (audiovisual) source separation exploiting multi-speaker tracking, robust beamforming and time-frequency masking. *IET Signal Processing*, 6(5):466–477. 20

NERC (2010). North american eletric reliability corporation (nerc), event analysis: System disturbance reports. *http://www.nerc.com*. 19

Ni, J., Qiu, Q., and Chellappa, R. (2013). Subspace interpolation via dictionary learning for unsupervised domain adaptation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 10, 14

Ophir, B., Lustig, M., and Elad, M. (2011). Multi-scale dictionary learning using wavelets. *IEEE Journal of Selected Topics in Signal Processing*, 5(5):1014–1025. 79

Paisitkriangkrai, S., Shen, C., and van den Hengel, A. (2015). Learning to rank in person re-identification with metric ensembles. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 15

Pan, S., Kwok, J., and Yang, Q. (2008). Transfer learning via dimensionality reduction. In *Conf. on AAAI*. 3

Pan, S., Tsang, I., Kwok, J., and Yang, Q. (2011). Domain adaptation via transfer component analysis. *IEEE Trans. on Neural Networks*, 22(3):199–210. 3

Pan, S. and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. on Knowledge and Data Engineering*, 22(3):13451359. 4

Paramesmaran, V. and Chellappa, R. (2006). View invariance for human action recognition. *Int. Journal on Computer Vision*, 66(1):83–101. 11

Pedagadi, S., Orwell, J., Velastin, S., and Boghossian, B. (2013). Local fisher discriminant analysisi for pedestrian re-identification. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 15, 52, 60, 67, 69

Phadke, A. and Thorp, J. (2008). *Synchronized Phasor Measurements and Their Applications*. New York: Springer. 18

Prosser, B., Zheng, W., Gong, S., Xiang, T., and Mary, Q. (2010). Person re-identification by support vector ranking. In *British Machine Vision Conference (BMVC)*. 15, 69, 70

Qi, H., Liu, Y., Li, F., and etc. (2011). Increasing the resolution of wide-area situational awareness of the power grid through event unmixing. In *IEEE 44th Hawaii International Conference on System Sciences (HICSS)*. 78

Qiu, Q., Patel, V., Turaga, P., and Chellappa, R. (2012). Domain adaptive dictionary learning. In *European Conference on Computer Vision (ECCV)*. 14

Quattoni, A., Collins, M., and Darrell, T. (2008). Transfer learning for image classification with sparse prototype representations. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 11

Ramirez, I., Sprechmann, P., and Sapiro, G. (2010). Classication and clustering via dictionary learning with structured incoherence. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 27

Rao, C., Yilmaz, A., and Shah, M. (2002). View invariance representation and recognition of actions. *Int. Journal on Computer Vision*, 50(2):203–226. 10

Remus, R. (2012). Domain adaptation using domain similarity- and domain complexity-based instance selection for cross-domain sentiment analysis. In *IEEE Int. Conf. on Data Mining Workshops*. 3

Ristin, M., Gall, J., Guillaumin, M., and Gool, L. V. (2015). From categories to subcategories: Large-scale image classification with partial class label refinement. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 17

Ristin, M., Guillaumin, M., Gall, J., and Gool, L. V. (2014). Incremental learning of ncm forests for large-scale image classification. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 17

Romero, D. G. and McCree, A. (2014). Supervised domain adaptation for i-vector based speaker recognition. In *IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP)*. 10

Schulter, S., Leistner, C., Wohlhart, P., Roth, P. M., and Bischof, H. (2013). Alternating regression forests for object detection and pose estimation. In *IEEE Int. Conf. on Computer Vision (ICCV)*. 17

Schulter, S., Leistner, C., Wohlhart, P., Roth, P. M., and Bischof, H. (2014). Accurate object detection with joint classification-regression random forests. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 17

Seo, Y., Lee, D., and Yoo, C. (2014). Salient object detection using bipartite dictionary. In *IEEE Int. Conf. on Image Processing (ICIP)*. 14

Shekhar, S., Patel, V. M., and Nguyen, H. V. (2013). Generalized domain-adaptive dictionaries. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 3

Shi, Z., Hospedales, T., and Xiang, T. (2015). Transferring a semantic representation for person re-identification and search. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 16

Shotton, J., Johnson, M., and Cipolla, R. (2008). Semantic texton forests for image categorization and segmentation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 17

Smeaton, A. and Over, P. (2003). Trecvid: Benchmarking the effectiveness of information retrieval tasks on digital video. In *Int. Conf. on Image and Video Retrieval*. 3

Song, Y., Wang, W., Zhang, Z., and Qi, H. (2015). Multiple event analysis for large-scale smart grid systems through cluster-based sparse coding. In *IEEE Int. Conf. on Smart Grid Communications*. 20

Sprechmann, P. and Sapiro, G. (2010a). Dictionary learning and sparse coding for unsupervised clustering. In *IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP)*, pages 2042–2045. 27

Sprechmann, P. and Sapiro, G. (2010b). Dictionary learning and sparse coding for unsupervised clustering. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 79

Taalimi, A., Ensafi, S., Qi, H., and Shijian Lu, Ashraf Kassim, C. L. T. (2015a). Multimodal dictionary learning and joint sparse representation for hep-2 cell classification. In *Int. Conf. on Medical Image Computing and Computer Assisted Interventions*. 14

Taalimi, A., Khorsandi, R., and Qi, H. (2015b). Online multi-modal task-driven dictionary learning and robust joint sparse representation for visual tracking. In *IEEE Conf. on Advanced Video and Signal Based Surveillance (AVSS)*. 14

Taixe, L. L., Fenzi, M., Kuznetsova, A., Rosenhahn, B., and Savarese, S. (2014). Learning an image-based motion context for multiple people tracking. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 17

Tan, D. J. and Ilic, S. (2014). Multi-forest tracker: A chameleon in tracking. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 17

Tang, X., Zhang, S., and Yao, H. (2013). Sparse coding based motion attention for abnormal event detection. In *IEEE Int. Conf. on Image Processing (ICIP)*. 14

Thorndike, E. and Woodworth, R. (1901). The influence of improvement in one mental function upon the efficiency of other functions: Functions involving attention, observation and discrimination. *Psychological Review*, 8(1):553564. 9

Thorp, J., Seyler, C., and Phadke, A. (1998). Electromechanical wave propagation in large electric power systems. *IEEE Trans. on Circuits System*, 45(6):614–622. 77

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58(1):267–288. 13, 83

Tran, C. and Trivedi, M. (2008). Human body modelling and tracking using volumetric representation: Selected recent studies and possibilities for extensions. ACM/IEEE Int. Conf. on Distributed Smart Cameras. 22

Tran, D. and Sorokin, A. (2008). human activity recognition with metric learning. In *European Conf. on Computer Vision (ECCV)*. 37

Trevor, H., Robert, T., and Jerome, F. (2008). *The Elements of Statistical Learning*. Springer. 16

Vapnik, V. (1998). *Statistical learning theory*. Wiley-Interscience. 2

Wang, J. and Zheng, H. (2012). Cross-view action recognition by statistical machine translation. *Biometric Recognition*, 77:60–67. 11

Wang, S., Zhang, L., Liang, Y., and Pan, Q. (2012). Semi-coupled dictionary learning with applications to image super-resolution. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 11, 23

Wang, W., Ayhan, B., Kwan, C., Qi, H., and Vance, S. (2013a). A novel and effective multivariate method for compositional analysis using laser induced breakdown spectral data. In *35th Int. Symposium on Remote Sensing of Environment (ISRSE)*. 19

Wang, W., He, L., Markham, P., Qi, H., and Liu, Y. (2013b). Detection, recognition, and localization of multiple attacks through event unmixing. In *IEEE Int. Conf. on Smart Grid Communication*. 14

Wang, W., He, L., Markham, P., Qi, H., Liu, Y., Cao, Q., and Tolbert, L. (2014a). Multiple event detection and recognition through sparse unmixing for high-resolution situational awareness in power grid. *IEEE Trans. on Smart Grid*, 5(4):1949–1664. 14

Wang, W., Li, S., Qi, H., Ayhan, B., Kwan, C., and Vance, S. (2014b). Revisiting the preprocessing procedures for the elemental concentration estimation based on chemcam libs on mars rover. In *IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (Whispers)*. 20

Wang, W., Li, S., Qi, H., Ayhan, B., Kwan, C., and Vance, S. (2015). Identify anomaly components in hyperspectral images by sparsity and low rank. In *IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (Whispers)*. 20

Wang, W., Liu, L., Zhan, L., Qi, H., and Liu, Y. (2013c). Highly accurate frequency estimation for fnet. In *IEEE Power Engineering Society General Meeting (PESGM)*. 19

Wang, W., Luo, J., and Qi, H. (2013d). Action recognition across cameras via reconstructable paths. In *IEEE Int. Conf. on Distributed Smart Cameras (ICDSC)*. 14

Wang, W. and Qi, H. (2013). Unsupervised non-linear unmixing of hyperspectral image based on sparsity constrained probabilistic latent semantic analysis. In *IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (Whispers)*. 19

Wang, W., Zhao, H., and Dong, C. (2009). A parallel algorithm of anomalies detection in hyperspectral image with projection pursuit - based on special projection searching. *Journal of BeiHang University*, 35(3):342–346. 20

Wang, X. (2013). Intelligent multi-camera video surveillance: A review. *Pattern Recognition Letter*, 34:3–19. 51

Weinland, D., Boyer, E., and Ron, R. (2007). Action recognition from arbitrary views using 3d exemplars. In *IEEE Int. Conf. on Computer Vision (ICCV)*. 11, 23, 35

Weinland, D., Ozuysal, M., and Fua, P. (2010). Making action recognition robust to occlusions and viewpoint changes. In *European Conf. on Computer Vision (ECCV)*. 11, 23

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers. 1

Wright, J., Yang, A., Ganesh, A., Sastry, S., and Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31(2):210–227. 13, 28, 82

Wu, D., Lee, W., Ye, N., and Chieu, H. L. (2009). Empirical study on the performance stability of named entity recognition model across domains. In *IEEE Conf. on Empirical Methods in Natural Language Processing*. 3

Wu, X. and Jia, Y. (2012). View-invariant action recognition using latent kernelized structural svm. In *European Conf. on Computer Vision (ECCV)*. 11, 23, 37, 45

Wu, Z., Li, Y., and Radke, R. (2015). Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 37(5):1095–1108. 15

Xia, T., Zhang, H., Gardner, R., Bank, J., Dong, J., Zuo, J., Liu, Y., Beard, L., Hirsch, P., Zhang, G., and Dong, R. (2007). Wide-area frequency based event location estimation. In *IEEE Power Engineering Society General Meeting (PESGM)*. 18, 73

Xiao, M. and Guo, Y. (2015). Feature space independent semi-supervised domain adaptation via kernel matching. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 37(1):54–66. 10

Xiong, F., Gou, M., Camps, O., and Sznaier, M. (2014). Person re-identification using kernel based metric learning methods. In *European Conf. on Computer Vision (ECCV)*. 15, 52, 67, 70

Xue, M. and Ling, H. (2009). Robust visual tracking using l1 minimization. In *IEEE Int. Conf. on Computer Vision (ICCV)*. 14

Yang, J., Wright, J., Huang, T., and Ma, Y. (2010a). Image super-resolution via sparse representation. *IEEE Trans. on Image Processing*, 9(11):2861–2873. 79

Yang, J., Yan, R., and Hauptmann, A. (2007). Cross-domain video concept detection using adaptive svms. In *ACM Conf. on Multimedia*. 9

Yang, J., Yu, K., and Huang, T. (2010b). Supervised translation-invariant sparse coding. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 13

Yang, Q., Chen, Y., Xue, G., Dai, W., and Yu, Y. (2009). Heterogeneous transfer learning for image clustering via the social web. In *IEEE Int. Joint Conf. on Natural Language Processing*. 3

Yi, D., Lei, Z., and Li, S. (2014). Deep metric learning for practical person re-identification. In *Int. Conf. on Pattern Recognition*. 16

Yilmaz, A. and Shah, M. (2005). Actions sketch: A novel action representation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 11

Yin, J., Yang, Q., and Ni, L. (2008). Learning adaptive temporal radio maps for signalstrength-based location estimation. *IEEE Trans. on Mobile Computing*, 7(3):869883. 3

Zhang, X. and Mahoor, M. (2014). Simultaneous detection of multiple facial action units via hierarchical task structure learning. In *Int. Conf. on Pattern Recognition*. 10

Zhang, Y., Markham, P., Xia, T., and et al (2010). Wide-area frequency monitoring network (fnet) architecture and applications. *IEEE Trans. on Smart Grid*, 1(2):159–167. 18, 73

Zhang, Z., Wang, C., Xiao, B., Zhou, W., Liu, S., and Shi, C. (2013). Cross view action recognition via a continuous virtual path. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. x, xiii, 12, 37, 38, 41, 42, 43, 44, 45

Zhao, Q., Dong, J., Xia, T., and Liu, Y. (2008). Detection of the start of frequency excursion in wide-area measurements. In *IEEE Power and Energy Society General Meeting (PESGM)*. 18, 19, 73, 78

Zhao, R., Ouyang, W., and Wang, X. (2013a). Person re-identification by salience matching. In *IEEE Int. Conf. on Computer Vision (ICCV)*. 14

Zhao, R., Ouyang, W., and Wang, X. (2013b). Unsupervised salience learning for person re-identification. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 14, 52, 60, 67, 69, 70

Zhao, R., Ouyang, W., and Wang, X. (2014). Learning mid-level filters for person re-identification. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 16, 61, 67, 69

Zheng, J. and Jiang, Z. (2013). Learning view-invariant and sparse representations for cross-view action recognition. In *IEEE Int. Conf. on Computer Vision (ICCV)*. 25, 26

Zheng, J., Jiang, Z., Phillips, P. J., and Chellappa, R. (2012). Cross-view action recognition via transferable dictionary pair. In *British Machine Vision Conference (BMVC)*. x, 12, 25, 26, 37, 41

Zheng, L., Wang, S., Tian, L., He, F., Liu, Z., and Tian, Q. (2015). Query adaptive late fusion for image search and person re-identification. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 67, 69

Zheng, W., Gong, S., and Xiang, T. (2011). Person re-identification by probabilistic relative distance comparison. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 15, 69, 70

Zheng, W., Gong, S., and Xiang, T. (2013). Reidentification by reative distance comparison. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(3):653–668. 52

Zhong, Z., Xu, C., Billian, B., Zhang, L., and et al (2005). Power system frequency monitoring network (fnet) implementation. *IEEE Trans. on Power Systems*, 20(4):1914–1921. 18, 73

Zhu, F. and Shao, L. (2014). Weakly supervised cross domain dictionary learning for visual recognition. *Int. Journal of Computer Vision*, 109(1):42–59. 14

Zhu, H. and Giannakis, G. (2012). Sparse overcomplete representations for efficient identification of power line outages. *IEEE Trans. on Power Systems*, 27(4):2215–2224. 19

Zhuang, F., Luo, P., Xiong, H., Xiong, Y., He, Q., and Shi, Z. (2010). Cross-domain learning from multiple sources: A consensus regularization perspective. *IEEE Trans. on Knowledge and data engineering*, 65(4):112–123. 3

Appendix

Publications

- **Computer Vision:**

1. **W. Wang**, A. Taalimi, K. Duan, R. Guo and H. Qi, "Learning Patch-Dependent Kernel Forest for Person Re-Identification", IEEE Winter Conference on Applications of Computer Vision, under review, 2016.
2. **W. Wang**, K. Duan, T.P. Tian, T. Yu and H. Qi, "Visual Tracking based on Object Appearance and Structure Preserved Local Patches Matching", in submission.
3. **W. Wang**, T. Gee, J. Price and H. Qi, "Real Time Multi-Vehicle Tracking and Counting at Intersections from a Fisheye Camera", IEEE Winter Conference on Applications of Computer Vision (WACV), 2015.
4. **W. Wang**, J. Luo and H. Qi, "Robust Cross View Action Recognition by Reconstructable Paths between View-Dependent Representations", IEEE Journal on Selected Topics in Signal Processing, 2015, under review.
5. **W. Wang**, J. Luo and H. Qi, "Human Action Recognition Across Cameras via Reconstructable Paths", ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC), 2013.
6. K. Duan, **W. Wang** and T. Yu, "Procrustean Decomposition for Orthogonal Cascade Detection", IEEE Winter Conference on Applications of Computer Vision (WACV), under review, 2016.
7. J. Luo, **W. Wang** and H. Qi, "Group Sparsity and Geometry Constrained Dictionary Learning for Action Recognition from Depth Maps", IEEE International Conference on Computer Vision (ICCV), 2013.
8. J. Luo, **W. Wang** and H. Qi, "Feature Extraction and Representation for Distributed Multi-View Human Action Recognition", IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 3(2):145-154, 2013.

- **Signal Pattern Recognition:**

  1. **W. Wang**, L. He, P. Markham, H. Qi, Y. Liu, Q. Cao and L. Tolbert, "Multiple Event Detection and Recognition through Sparse Unmixing for High-Resolution Situational Awareness in Smart Grid", IEEE Transaction on Smart Grid, 2014.
  2. **W. Wang**, L. He, P. Markham, H. Qi and Y. Liu, "Detection, Recognition, and Localization of Multiple Attacks through Event Unmixing", IEEE International Conference on Smart Grid Communications, 2013.
  3. **W. Wang**, L. Liu, L. He, H. Qi and Y. Liu, "Highly Accurate Frequency Estimation for FNET", IEEE PES General Meeting (PESGM), 2013.
  4. Y. Song, **W. Wang**, Z. Zhang, H. Qi and Y. Liu, "Multiple Event Analysis for Large-scale Smart Grid Systems Through Cluster-based Sparse Coding", IEEE International Conf. on Smart Grid Communications, 2015.
  5. Z. Zhang, Y. Song, **W. Wang** and H. Qi, "On-line Modeling and Classification of Streaming Time Series Using Derivative Delay Embedding", submitted to AAAI, under review, 2016.

- **Hyperspectral Image Analysis**

  1. **W. Wang**, S. Li, H. Qi, B. Ayhan, C. Kwan and S. Vance, "Identify Anomaly Component by Sparsity and Low Rank Analysis", IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, (WHISPERS), 2015.
  2. **W. Wang**, S. Li, H. Qi, B. Ayhan, C. Kwan and S. Vance, "Revisiting the Preprocessing Procedures for Elemental Concentration Estimation based on ChemCam LIBS on MARS Rover", IEEE WHISPERS, 2014.
  3. **W. Wang** and H. Qi, "Unsupervised Non-linear Unmixing of Hyperspectral Image based on Sparsity Constrained Probabilistic Latent Semantic Analysis", IEEE WHISPERS, 2013.

4. **W. Wang**, B. Ayhan, C. Kwan, H. Qi, "A Novel and Effective Multivariate Method for Compositional Analysis using Laser Induced Breakdown Spectroscopy", 35th Int. Symp. on Remote Sensing of Environment, 2013.
5. R. Guo, **W. Wang** and H. Qi, "Hyperspectral Image Unmixing using Cascaded AutoEncoder", IEEE WHISPERS, 2015.
6. S. Li, **W. Wang**, H. Qi, B. Ayhan, C. Kwan and S. Vance, "Low-rank tensor decomposition based anomaly detection for hyperspectral imagery", IEEE International Conference on Image Processing (ICIP), 2015.

Vita

Wei Wang was born in Hanshou county, Changde city, Hunan province, P.R. China. He received his Bachelor and Master degrees in Electrical Engineering from University of Science and Technology Beijing (USTB) and BeiHang University (BUAA) in 2006 and 2009, respectively. Then, he worked as an assistant engineer in Beijing Railway Research and Development Institute of Signal and Communication. From fall 2010, he enrolled into the doctoral program at the University of Tennessee at Knoxville in the department of Electrical Engineering and Computer Science. At the same time, he joined the Advanced Imaging and Collaborative Information Processing (AICIP) lab as a graduate research assistant under supervision of Professor Hairong Qi. His major research areas are: computer vision, pattern recognition, signal processing and parallel computing.